h e g

Haute école de gestion
Genève

L'archivage du web :

stratégies, études de cas et recommandations

Travail de Bachelor réalisé en vue de l'obtention du Bachelor HES

par :

Jonas BEAUSIRE

Conseiller au travail de Bachelor :

Françoise DUBOSSON NALO, chargée d'enseignement

Genève, 13 juillet 2015

Haute École de Gestion de Genève (HEG-GE)

Filière Information documentaire

Hes·so GENÈVE
Haute Ecole Spécialisée
de Suisse occidentale

Déclaration

Ce travail de Bachelor est réalisé dans le cadre de l'examen final de la Haute école de gestion de Genève, en vue de l'obtention du titre Bachelor of Science HES-SO en Information documentaire.

L'étudiant atteste que son travail a été vérifié par un logiciel de détection de plagiat.

L'étudiant accepte, le cas échéant, la clause de confidentialité. L'utilisation des conclusions et recommandations formulées dans le travail de Bachelor, sans préjuger de leur valeur, n'engage ni la responsabilité de l'auteur, ni celle du conseiller au travail de Bachelor, du juré et de la HEG.

« J'atteste avoir réalisé seul le présent travail, sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Lausanne, le 10 juillet 2015

Jonas Beausire

Remerciements

J’aimerais remercier en premier lieu Françoise Dubosson, ma conseillère, qui a su, tout au long de mon travail, m’accompagner avec intelligence et gentillesse. Je tiens également à remercier Brigitte Steudler et Annick Le Follic pour le temps précieux qu’elles m’ont toutes deux accordé : leur expertise m’a été d’une grande aide. Je remercie Enrico Natale pour les sources très utiles qu’il a eu l’amabilité de partager avec moi.

Je remercie aussi chaleureusement Monique Beausire, Alenka Bonnard et Alexandre Dayer, mes relecteurs, dont le regard et l’acuité ont permis la naissance de ce travail.

Enfin, je remercie tous ceux qui, dans mon entourage, ont accompagné ce mémoire : Guillaume Beausire, Philippe Blatti, Benoît Bovay et Olivier Dorsaz.

Résumé

Ce travail consiste en l'établissement d'un panorama des grandes approches et stratégies de collecte de l'archivage du web, une analyse des attentes et des résistances du public des chercheurs face à ces nouvelles archives et la présentation de pistes d'innovations et de recommandations pour mieux appréhender l'archivage du web. Une analyse approfondie de deux programmes d'archivage – celui de la Bibliothèque nationale suisse (BN) et celui de la Bibliothèque nationale de France (BnF) – et une comparaison de ces deux modèles le complètent.

Une revue générale, puis spécifique, de la littérature consacrée à l'archivage du web a été nécessaire. Les sources proviennent toutes de bases de données et du web. Des entretiens exploratoires qualitatifs semi-directifs ont été menés afin d'éclairer les points d'ombre des sources préalablement investies. Un travail de synthèse et de compilation de l'ensemble des sources et des entretiens a mené à la rédaction de ce travail.

Les approches de l'archivage du web sont exposées : intégrale, exhaustive, sélective et thématique. Elles se combinent souvent sur le terrain mais doivent être repensées pour être renouvelées. Chacune d'entre-elles peut être accompagnée d'une stratégie de collecte : automatisée, semi-automatisée ou manuelle. Les logiques juridiques et patrimoniales, ainsi que les processus de travail des programmes d'archivage de la BN et de la BnF sont détaillés : l'arsenal juridique structure fondamentalement les possibilités des deux institutions. Les attentes des chercheurs, leurs besoins et résistances sont mis en lumière par des résultats d'enquêtes. Si la communauté scientifique s'accorde sur la nécessité de constituer une mémoire du web, la fiabilité et la légitimité des collections issues du web cristallisent les résistances exprimées par les chercheurs. Globalement, les questions épistémologiques et méthodologiques pour inscrire ces archives dans un usage scientifique établi ne sont pas encore résolues. Enfin, des recommandations techniques et conceptuelles sont abordées : elles mettent notamment l'accent sur la construction d'interfaces d'accès et la description des archives et de leur contexte grâce, en particulier, aux métadonnées. Une variété d'outils d'analyse du web constitue également des leviers privilégiés pour exploiter et mettre en valeur les futures archives du web.

Table des matières

Liste des tableaux

1. Introduction

Les questions soulevées par l'archivage du web préoccupent les acteurs du monde de l'information et des archives depuis presque vingt ans maintenant. Des initiatives comme celles de la fondation « Internet Archive » ou de la Bibliothèque nationale de Suède ont pris naissance dès 1996. La mise en place de principes fondateurs et les premières expérimentations des méthodes d'archivage du web ont ainsi vu le jour. Les institutions concernées ont immédiatement pointé un double constat : d'une part, la production éditoriale née numérique possède une valeur patrimoniale : « Le web, à la fois par le nombre et la variété des contenus qu'il met à disposition, […] est […] devenu une part majeure de notre patrimoine. » (Bonnel, Oury, 2014, p. 2), d'autre part, l'indubitable disparition du web d'hier est toujours plus importante.

Les pertes, très tôt constatées par les administrateurs des programmes d'archivage du web, sont la conséquence directe d'une très grande fragilité des documents issus de l'Internet. Tout au long de ce travail, nous n'aurons de cesse de souligner les dimensions éphémères, fuyantes et nomades (Genin 2012, p. 21) des contenus présents sur le web. L'urgence de leurs collectes s'est peu à peu répandue au sein de grandes institutions patrimoniales et des cadres législatifs ont vu le jour pour s'emparer au mieux de ces documents, symptômes d'une « accréditation culturelle de l'éphémère » (Merzeau 2003, p. 1). Les enjeux de la sauvegarde de cette mémoire numérique inquiètent même jusqu'aux sphères les plus dominantes, puisque le vice-président de Google, Vinton Cerf, a récemment lancé un appel alarmiste : « When you think about quantity of documentation from our daily lives that is captured in digital form, […], it's clear that we stand to lose an awful lot of our history. » (Sample 2015). Perpétuant les buts traditionnels des archives « classiques », les archives du web conservent ainsi leurs fonctions de préservation, d'authentification et de mise à disposition. Néanmoins, la constitution de ces nouvelles collections d'archives n'est pas sans poser plusieurs questions qui se retrouveront au cœur de ce travail : selon quelle approche théorique peut-on se saisir de ces documents ? Comment travaillent les institutions chargées de la collecte des documents du web ? A quel public ces archives se destinent-elles ? Vingt ans après les premières initiatives, quelles sont les perspectives et innovations futures de cet archivage particulier ?

Afin de tenter de répondre aux questions énumérées plus haut, ce travail s'ouvrira, suite à cette introduction et à notre méthodologie, sur une présentation des grandes approches et stratégies de collecte de l'archivage du web. Ce sera l'occasion de dresser un panorama théorique général des processus d'archivage à l'œuvre et de

situer deux études de cas développées au chapitre quatre. Les limites de ces approches seront abordées et des exemples du terrain viendront illustrer chacune d'entre elles.

Nous nous pencherons ensuite sur l'analyse de deux programmes d'archivage du web : celui de la Bibliothèque nationale suisse (BN), « Archives Web Suisse » et celui de la Bibliothèque nationale de France (BnF), « Archives de l'Internet ». Une étude approfondie des deux programmes et une comparaison de ces deux modèles à l'œuvre composeront le chapitre quatre de ce travail. L'analyse spécifique du cadre législatif, technique et archivistique de chacun des deux programmes permettra de saisir les réalités du terrain auxquelles sont confrontés les professionnels. Ce chapitre sera enrichi d'entretiens avec certains responsables qui rapporteront leur expertise et leurs expériences.

Nous aborderons, au chapitre cinq, la question du public de ces nouvelles collections issues des différents programmes d'archivage du web. En effet, à quels segments cette mémoire patrimoniale du numérique s'adresse-t-elle ? Parmi la variété des publics possibles, nous nous pencherons spécifiquement sur celui des chercheurs et des universitaires. Les besoins, les attentes et les résistances de cette population face à ces nouvelles sources seront abordés, notamment grâce aux résultats de certaines enquêtes. Nous mettrons également en lumière les communautés scientifiques les plus concernées par la mobilisation de ces archives et la force des collaborations entre chercheurs et acteurs des programmes.

Enfin, le dernier chapitre de ce travail présentera un panorama non-exhaustif des futurs possibles de l'archivage du web. Les pistes d'innovations sont nombreuses et entreront parfois en écho avec les programmes étudiés ou les besoins des chercheurs exposés aux chapitres précédents. Les inspirations pour une meilleure exploitation des archives proviennent souvent d'outils d'analyse du web vivant. L'ensemble de ces pistes pourra se lire comme un faisceau de recommandations variées pour mieux penser et valoriser le travail des responsables des programmes de ces archives très particulières.

Nous espérons ainsi atteindre nos objectifs qui sont la présentation d'un panorama des grandes approches de l'archivage du web existantes, une analyse des attentes et des résistances du public des chercheurs face à ces nouvelles archives et l'établissement de pistes et de recommandations pour mieux appréhender l'archivage du web.

Les enjeux de l'archivage du web sont multiples ; au-delà des nombreuses questions techniques et organisationnelles, se posent celles, plus réflexives, liées au patrimoine et à la mémoire de nos sociétés. En organisant la masse gigantesque de contenus désormais dématérialisés du web, les institutions mandataires opèrent des choix, sélectionnent et architecturent notre mémoire collective de demain. Cette responsabilité immense soulève des questions éminemment politiques : si Internet est aujourd'hui un espace qui appartient à tout le monde (Illien 2011), comment le sauvegarder équitablement ? Accessibilité, représentativité, légitimité, fiabilité et destruction des archives du web sont quelques-unes des préoccupations récurrentes qui jalonneront ce travail.

2. Méthodologie

Afin de mener à bien ce travail, nous avons en tout premier lieu établi une revue de la littérature consacrée à l'archivage du web. Nous avons commencé par nous procurer des sources générales pour mieux comprendre les problématiques et les enjeux au cœur de ce nouvel archivage. Comme la problématique se cristallisait initialement autour des documents audiovisuels numériques natifs, nous avons ensuite orienté nos recherches vers des sources plus spécifiques. Nous avons rapidement constaté que très peu de sources traitaient précisément de l'archivage de ce type de document. Comme nous l'explicitons au début du chapitre quatre, nous avons dû renoncer, en cours de route, à utiliser la focale du document audiovisuel pour l'analyse des deux études de cas. L'objet-test devenant caduc, nous avons élargi nos recherches en nous concentrant désormais sur les documents traitant de l'archivage au sein de la BN et de la BnF, de la question du public des chercheurs et enfin des innovations technologiques et des défis futurs. Nous avons ainsi sollicité plusieurs bases de données, principalement LISA (Library and Information science abstracts), LISTA (Library, Information Science and Technology Abstracts) et Cairn. Comprenant rapidement qu'une littérature importante circulait au sein d'un consortium international, l'IIPC (International Internet Preservation Consortium), nous avons également procédé à des recherches actives sur Internet qui s'est révélé être le lieu le plus fécond en termes de sources. Nous avons également, mais dans une moindre mesure, consulté certaines banques de données de presse (Nexis et Factiva) au début de nos recherches : cela a été utile pour mieux cerner le sujet et s'informer des dernières actualités.

Lors de la lecture de ces sources, nous avons rapidement observé certaines lacunes, notamment concernant les deux programmes d'archivage analysés. Dans ce contexte, nous avons mené plusieurs entretiens exploratoires qualitatifs semi-directifs afin d'éclairer les points d'ombre des sources préalablement investies. Brigitte Steudler, responsable de la Documentation vaudoise au sein de la Bibliothèque cantonale et universitaire de Lausanne (BCU) et personne de contact opérationnel dans le cadre d'Archives Web Suisse pour le canton de Vaud, a été la première personne à nous recevoir. L'entretien a été enregistré puis synthétisé au sein d'un sous-chapitre sous la forme du discours rapporté. Nous avons choisi de ne pas retranscrire cet entretien et d'en offrir une synthèse fidèle et complète sous une forme plus rédigée. Le travail de terrain effectué par Brigitte Steudler en faisait une personne ressource extrêmement précieuse pour comprendre les processus de travail à l'œuvre dans le programme

Archives Web Suisse. La documentation, très abondante, sur le programme de la BN n'éclairait pas bon nombre de subtilités qui ont pu être récoltées lors de cet entretien. En ce qui concerne l'entretien téléphonique avec Annick Le Follic, responsable des collectes pour le dépôt légal numérique à la BnF, nous avons adopté exactement la même méthode : un sous-chapitre est consacré à une description complète de l'échange. En contactant Gildas Illien, directeur du Département Information bibliographique et numérique de la BnF, nous avons été redirigé auprès d'Annick Le Follic, plus à même de répondre à la grille d'entretien que nous avions jointe. Il est à noter que nous avons cherché, notamment, à communiquer par « Skype », mais que pour des raisons logistiques, nous nous sommes résolus à un entretien téléphonique qui a été néanmoins enregistré. L'échange de courriels avec Barbara Signori, responsable du programme e-Helvetica de la BN, n'a malheureusement pas pu déboucher sur un entretien formel, même au prix de certaines contorsions. Nous nous sommes finalement résolus à transmettre ce questionnaire par courriel à la responsable qui nous a renvoyé ses réponses par écrit. Cet échange a été mis en annexe de ce travail et constitue ainsi une source à part entière. Il nous a en effet semblé plus simple, au vu de la brièveté de ses réponses, de retranscrire fidèlement les quelques fragments obtenus. La structure de la grille d'entretien utilisée a été adaptée systématiquement en fonction de la personne interviewée. Néanmoins, une organisation thématique des questions s'y retrouvait invariablement :

- Processus de travail
- Périmètre de la collecte
- Gestion des documents audiovisuels
- BN et BnF : quel regard portent-elles l'une sur l'autre ?
- L'institution sur le plan international : quelle collaboration ?
- Futurs et défis de l'institution

Dans la perspective de la problématique initiale activant les documents pornographiques numériques natifs, nous avons rencontré deux universitaires pour des entretiens exploratoires : Thierry Delessert, chargé de cours, chercheur FNS senior 2^{e} année à l'UNIL, et Gary Crosilla, doctorant en sociologie. Ces deux entretiens n'ont pas pu être utilisés pour les raisons évoquées plus haut.

Enfin, après avoir remanié le plan initialement prévu, nous avons procédé à un important travail de synthèse et de compilation de l'ensemble des sources et des entretiens. Ces derniers ont été d'une aide capitale pour l'analyse des deux programmes et la rédaction du chapitre concernant le public des chercheurs. A chaque fois que cela était possible, nous avons essayé de mobiliser les sources les plus

récentes. En effet, les innovations techniques très rapides ont rendu certaines sources difficiles à utiliser. Cette synthèse a permis par la suite la rédaction de ce travail.

3. Grandes approches de l'archivage du web et stratégies de collectes[1]

« La grosse erreur, ce serait de considérer que le web c'est l'équivalent des paroisses du Moyen-Âge et que je vais récolter tout ça. C'est un processus dynamique qui continue toujours. Or, en essayant d'éliminer la dimension temporelle, on va être largement à côté. »

Un chercheur interrogé sur ses appréhensions des collections issues de l'archivage du web de la BnF (Chevallier, Illien, 2011, p. 14)

Ce chapitre vise à dresser un panorama très général et introductif des grandes approches et stratégies de l'archivage du web. Ce premier tour d'horizon sera l'occasion de situer brièvement les deux études de cas de notre travail dans le contexte mondial des archives du web. Afin d'offrir une typologie des grandes *approches* à l'œuvre au sein des différents programmes d'archivage dans le monde, nous allons reprendre celle conceptualisée par Thomas Chaimbault, responsable de la formation des bibliothécaires à l'ENSSIB. En effet, ce dernier déploie un panorama de stratégies et de modes de dépôt, développés par différents établissements nationaux et soutenus par des consortia. Nous exposerons également les trois *stratégies* de collectes qu'il propose au sein de son dossier documentaire (Chaimbault 2008).

Quatre grandes approches de collectes sont décrites : l'approche intégrale, l'approche exhaustive, l'approche sélective et l'approche thématique. Aucune d'entre elles n'apparaît comme parfaitement satisfaisante et des approches combinées sont donc souvent à l'œuvre, comme au sein des deux programmes étudiés au chapitre suivant. Chacune des approches est accompagnée d'un exemple du terrain qui vient illustrer brièvement les rouages de son application. Elles se combinent à des stratégies de collectes : automatiques, semi-automatiques ou manuelles. Ces stratégies renvoient aux ressources nécessaires pour déployer un projet d'archivage du web.

Le renouvellement des modèles d'archivage est au centre des préoccupations des professionnels, notamment en raison du dynamisme généralisé des documents présents sur le web : « Il ne faut […] pas tenter de transposer d'anciens modèles d'archivage. Il faut plutôt aller chercher du côté d'autres pratiques scientifiques des modèles nouveaux permettant de parler de cette archive absolument singulière. »

[1] L'entier de ce chapitre repose sur le travail de Thomas Chaimbault (Chaimbault 2008). Il est à noter que les grandes approches et stratégies exposées par Thomas Chaimbault dans son travail ainsi que les exemples qui les accompagnent ont été repris de l'article publié par Mehdi Gharsallah (Gharsallah 2004).

(Chevallier, Illien, 2011, p. 14). Ces grandes approches doivent ainsi sans cesse être questionnées et en aucun cas considérées comme gravées dans le marbre.

3.1 L'approche intégrale

Comme son nom l'indique, l'approche intégrale consiste à collecter l'entier du web, sans distinction ni critère de sélection. Les éventuelles valeurs patrimoniales ou documentaires sont évacuées au profit d'un projet chimérique d'exhaustivité. Le projet « Internet Archive »[2] en est l'exemple unique et donc le plus probant. Engagée en mars 1996, il s'agit aujourd'hui de « la plus importante archive du web » avec près de 480 milliards de pages archivées[3] et une croissance mensuelle vertigineuse de 100 téraoctets (Bnf 2014b). Cette bibliothèque virtuelle se découpe en cinq volets : la « Wayback machine », les images animées, les textes, les documents sonores et les logiciels. Même si les collections de documents audiovisuels augmentent considérablement depuis quelques années, l'onglet web (« Wayback machine ») qui archive les sites publics et permet « d'en retrouver les évolutions au fil du temps » (Peyssard 2012) demeure la collection la plus importante et la plus connue du projet.

Si ce projet aspire le web dans sa globalité, la fondation s'engage également dans des projets de collections circonscrites à des thématiques ciblées qui s'approchent d'une entreprise plus sélective ou thématique (voir plus bas) de l'archivage du web. C'est le cas de collections concernant, par exemple, les élections fédérales américaines.

3.2 L'approche exhaustive

A l'instar de l'approche intégrale, ce type d'archivage du web vise également une certaine idée de l'exhaustivité mais dans un périmètre précis, celui d'un nom de domaine, d'un espace national particulier ou, moins souvent, d'un type de sites. Il s'agit d'une appréhension de la collecte du web relativement répandue puisqu'elle peut s'inscrire facilement dans les missions d'une institution patrimoniale comme celles des bibliothèques nationales. Néanmoins, nous avons observé précédemment les ambiguïtés liées à la territorialité du web : des contenus web particulièrement signifiants peuvent être enregistrés sous un nom de domaine hors collecte, par exemple. La volatilité intrinsèque de l'Internet peut venir contrecarrer des projets d'exhaustivité comme celui mené par « Kulturarw3 »[4] en Suède. Dans le cadre de ce

2 Pour davantage d'informations sur le projet et pour notamment accéder à la « Wayback machine », consulter : https://archive.org/index.php
3 Chiffre au 1er juin 2015 disponible sur le site : http://www.archive.org/
4 Pour davantage d'informations sur ce programme d'archivage, consulter : http://www.kb.se/om/projekt/Svenska-webbsidor---Kulturarw3/

projet, la bibliothèque royale de Suède s'est proposée de récolter l'ensemble du domaine *.se*, visant à offrir les ressources documentaires les plus larges possibles aux chercheurs, sans préjuger de leurs futures attentes. Il s'agit bien ici d'une approche exhaustive et territoriale. Néanmoins, le projet semble souffrir d'une indexation peu efficace d'une masse documentaire particulièrement hétérogène. Nous verrons plus loin dans notre travail comment la BnF appréhende son exigence d'exhaustivité et selon quelles modalités.

3.3 L'approche sélective

A l'inverse de l'approche intégrale ou exhaustive, cette approche de l'archivage du web consiste précisément à se saisir de certains contenus prédéfinis au moyen de critères choisis. Ceux-ci peuvent être extrêmement variés : thématiques, en lien avec la nature de la ressource, qualitatifs, etc. Cette approche qui rompt avec un certain souci d'exhaustivité cherche à compiler régulièrement des instantanés de sites. L'archivage pratiqué par la BN s'inscrit parfaitement dans cette approche ; la liste des critères de sélection définit le périmètre de la collecte. Il est à noter que la BN pratique également un archivage dit « thématique » comme explicité plus bas.

Le projet de la bibliothèque nationale australienne « PANDORA »[5] participe également d'une telle approche. Lancé en 1996 en collaboration avec les Archives nationales, il vise à récolter un périmètre particulier de sites qui ont été préalablement sélectionnés, ainsi que des périodiques électroniques et des publications officielles. Cette approche sélective applique des critères de pertinence et de qualité aux ressources choisies. Le projet implique un panorama varié de partenaires (bibliothèques spécialisées, centres de recherche, etc.) : chacun est invité à choisir, décrire et traiter les sites sélectionnés. Nous retrouvons au travers de cet exemple australien les grandes diagonales qui traversent le projet de la BN que nous expliciterons en détails dans ce travail.

3.4 L'approche thématique

Cette dernière approche doit se comprendre comme un embranchement particulier de l'approche sélective : il s'agit ici d'archiver une collection de site web en lien avec un événement spécifique. Nous observerons, tout au long de ce travail, différents exemples de cette approche, notamment au travers des « collectes projet » menées par la BnF et certains moissonnages menés par la BN. Les collectes des sites web et autres ressources des élections présidentielles françaises en sont le parfait exemple (voir l'introduction du chapitre cinq (5.1)). Cette approche, tout comme l'approche

[5] Pour davantage d'informations sur ce programme d'archivage, consulter : http://pandora.nla.gov.au/

sélective, renvoie directement à la notion de collection, voire de « fonds d'archive » puisqu'il s'agit bien pour les bibliothécaires et les partenaires de sélectionner et d'éliminer en vue de former un corpus cohérent. Dans cette perspective, il est à noter que les collections ainsi formées peuvent représenter de véritables « produits d'appel » (Illien 2008) pour les utilisateurs. C'est en effet probablement la meilleure façon de sensibiliser un (nouveau) public aux richesses de ces nouvelles collections.

Force est de constater que les modèles conceptuels semblent souvent insuffisants pour encadrer les multiples réalités du terrain. Dès lors, ce sont des approches combinées qui sont privilégiées, comme nous le verrons au travers de nos deux études de cas au chapitre suivant.

3.5 Stratégies de récolte

Parallèlement aux différentes approches générales de l'archivage du web décrites plus haut, Thomas Chaimbault présente trois stratégies de collectes différentes : les stratégies automatisée, semi-automatisée et manuelle. La stratégie automatisée engage la mise en place d'un logiciel-robot comme nous pourrons l'observer à la fois au sein de la BN et de la BnF : un espace web circonscrit à un domaine choisi est ainsi collecté de façon automatique. Cette stratégie accompagne généralement des approches intégrales ou exhaustives de l'archivage du web. La stratégie de collecte semi-automatisée implique également l'usage d'un logiciel-robot mais ajoute à son utilisation des critères de sélection plus précis ; elle peut être mobilisée dans le cadre d'une approche sélective du web. Enfin, l'approche manuelle, même si elle exige également des ressources techniques, replace l'humain au centre des processus de collecte. Cette logique combinatoire est essentielle dans le contexte d'une approche thématique, par exemple. Les bibliothécaires sont ainsi amenés à sélectionner eux-mêmes les sites pertinents, ainsi qu'à « […] identifier, sélectionner, et collecter les sites du web profond […]. » (Chaimbault 2008).

3.6 Conclusion et récapitulatif

En conclusion de ce chapitre, nous présentons ici un tableau récapitulatif des grandes approches et stratégies développées plus haut. Nous rendons le lecteur attentif à la dimension réductrice d'un tel effort de synthèse. En effet, comme nous le verrons au chapitre suivant au travers des deux études de cas, les réalités du terrain sont multiples et mêlent bien souvent plusieurs approches et stratégies. Dans cette perspective, le tableau présente un récapitulatif caricatural : les cases « Néant » pourraient être parfois remplies, notamment celle de l'approche thématique semi-automatisée. Nous avons ainsi choisi de retenir ici uniquement les « cas d'école ».

Enfin, nous rappelons ici également toutes les limites théoriques de ces modèles à repenser en permanence, notamment en raison des mutations techniques extrêmement rapides du web. Plus encore, la nécessité d'inventer à l'avenir de nouveaux modèles qui ne reposent pas sur d'anciennes traditions de l'archivage apparaît comme cardinale. La nature tentaculaire du réseau, le dynamisme des documents et leur durée de vie invitent à repenser en permanence les modèles. Comme le propose l'un des chercheurs interrogé dans une étude sur les représentations et les attentes des chercheurs face aux collections du web archivé, il s'agirait davantage d'appréhender le web dans son flux dynamique, comme une « archive orale », plutôt que comme une archive silencieuse et figée : « Il faut que les gens du livre oublient un peu leur tradition documentaire et se disent : « On est dans les sociétés de tradition orale » ». Comment archiver un flux, plutôt que des unités documentaires isolées ? (Chevallier, Illien, 2011, p. 14). Ainsi, c'est plus spécifiquement la question de l'unité documentaire qui est bouleversée par ces archives d'un genre nouveau. Soulignons, en guise de conclusion et avant l'étude approfondie des deux programmes d'archivage, l'avertissement de Claude Mussou qui invite à la remise en question :

> *[…] si le site peut faire l'objet d'une description documentaire, des œuvres telles que les web documentaires, les web fictions, ou encore les pages profils des utilisateurs de réseaux sociaux sont autant de ces objets d'un nouveau type qui imposent de redéfinir l'approche documentaire[6]. »* *(2012)*

[6] C'est moi qui souligne.

Tableau 1 : Récapitulatif des grandes approches et stratégies

	Stratégie automatisée	**Stratégie semi-automatisée**	**Stratégie manuelle**
Approche intégrale	• Entier du web • Pas de critère de sélection • Logiciel-robot • **Ex : Internet Archive**	Néant	Néant
Approche exhaustive	• Entier du web, mais périmètre précis • Nom de domaine ou espace national • Logiciel-robot • **Ex : Kulturarw3**	Néant	Néant
Approche sélective	Néant	• Critères de sélection précis • Ressources humaines • Logiciel-robot • **Ex : BN ou PANDORA**	Néant
Approche thématique	Néant	Néant	• Critères de sélection précis • Collecte événementielle / thématique • Ressources humaines • **Ex : Collecte-projet de la BnF**

4. Etudes de cas des programmes d'archivage du web de la Bibliothèque nationale suisse (BN) et de la Bibliothèque nationale de France (BnF)

« Même si les gens qui publient n'en sont pas forcément conscients, on ne pourrait pas dans quelques années, ne serait-ce que dans dix ans, faire l'histoire du début du XXIe siècle, comprendre la société, si on n'a pas gardé la trace de ce qu'était le web en 2013. »

Clément Oury, dans un article de France Info (Beaudoux 2013)

Après avoir présenté le panorama des grandes approches de l'archivage du web dans le monde et explicité les stratégies de collectes en vigueur, nous allons resserrer notre travail autour de deux études de cas. En effet, nous allons procéder à une analyse fine de deux programmes d'archivage du web spécifiques : celui de la Bibliothèque nationale suisse (BN), « Archives Web Suisse » et celui de la Bibliothèque nationale de France (BnF), « Archives de l'Internet ». Après l'étude des programmes respectifs des deux institutions, nous procéderons à une comparaison de ces deux modèles. Issus d'approches très différentes mais néanmoins conceptualisés au sein d'institutions ayant des vocations similaires – les deux structures responsables sont des bibliothèques nationales patrimoniales – les deux programmes étudiés donneront à voir un cadre législatif, technique et archivistique très différent. Nous verrons notamment comment le cadre juridique structure les logiques de collecte des documents numériques et les conditions d'accès aux collections. Nous observerons également, lors de la comparaison des programmes, la façon dont ces deux approches peuvent se compléter et se répondre, notamment dans le cadre de collaborations internationales.

La question de l'archivage des documents audiovisuels numériques natifs, et plus spécifiquement celle des documents pornographiques comme outil d'évaluation (« objet-test ») des politiques d'archivage, explicitée au sein du cahier des charges de ce travail, n'a pu être posée, ni résolue dans l'étude des deux programmes. Nous avons rapidement constaté que les deux programmes étudiés ne considéraient pas l'archivage du web à un tel degré de granularité. En effet, c'est bien plutôt l'écosystème du site web dans son entier qui est analysé par les différents acteurs des programmes et non les publications internes au site. En mobilisant une catégorie spécifique de documents pour analyser ces deux politiques d'archivage, nous avons anticipé sans le savoir les perspectives de « redocumentarisation », conceptualisées par Jean-Michel Salaün. Ce nouveau paradigme consiste ni plus ni moins à documentariser des

ressources numériques, c'est à dire : « […] traiter un document comme le font […] les professionnels de la documentation […] : le cataloguer, l'indexer, le résumer, le découper, éventuellement le renforcer, etc. », avec pour objectif final « […] d'optimiser l'usage du document en permettant un meilleur accès à son contenu et une meilleure mise en contexte. » (Salaün 2007). Or, la plasticité propre aux contenus numériques natifs bouleverse le geste et les pratiques de documentarisation acquises jusqu'ici par les professionnels de l'information. En effet, comment documentariser des contenus sans cesse enrichis et mouvants, désormais inscrits sur des supports eux-mêmes instables (obsolescence des formats de fichier) : « […] bien des unités documentaires du Web ne ressemblent plus que de très loin aux documents traditionnels. […] la stabilité du document classique s'estompe et la redocumentarisation prend une toute autre dimension. » (Salaün 2007). Les solutions avancées par Jean-Michel Salaün et d'autres chercheurs se cristallisent aujourd'hui autour de l'apport des métadonnées : « Ces [nouveaux] différents niveaux d'accès nécessitent que soient créées des métadonnées de même niveau de granularité. » (Han, 2012, p. 1). En effet, une application concrète et systématique d'un jeu de métadonnées à de très gros volumes de contenus constitue un enjeu majeur du travail de redocumentarisation. Le document audiovisuel numérique natif (pornographique) comme objet-test de notre évaluation aurait pu fonctionner si les réflexes de redocumentarisation décrits plus haut étaient déjà inscrits au cœur de l'appréhension de la masse du web par les différents acteurs de son archivage.

Ainsi, même si l'étude des deux programmes d'archivage de ce chapitre a révélé cette réalité et propose donc une analyse plus globale, sans la focale audiovisuelle préalablement choisie, il n'en demeure pas moins que les grandes diagonales à la fois juridiques, patrimoniales et pratiques ont pu être mises en lumière.

4.1 BN : projet e-Helvetica

Le projet e-Helvetica engagé par la Bibliothèque nationale suisse (BN) depuis 2001 s'inscrit dans l'une des missions fondamentales des bibliothèques qui est la sauvegarde des documents désormais indisponibles ou risquant de le devenir. L'usage massif d'Internet tend à considérer les documents y circulant comme de véritables « objets de la mémoire » (Balzardi 2008). Dans cette perspective, la BN et son projet e-Helvetica ont pour objectifs principaux la mise en place des bases de collecte, de saisie, d'archivage et de mise à disposition des Helvetica électroniques (offline ou online) (Balzardi 2008), ainsi que la constitution de collections d'Helvetica numériques.

La notion d'Helvetica est réglée juridiquement à l'article 3 de la Loi fédérale sur la Bibliothèque nationale suisse (LBNS). Celui-ci définit l'Helvetica comme :

> *« […] les informations imprimées ou conservées sur d'autres supports que le papier qui paraissent en Suisse, se rapportant à la Suisse, à ses ressortissants ou à ses habitants ou sont créés, en partie ou en totalité, par des auteurs suisses ou par des auteurs étrangers liés à la Suisse. » (Suisse 1992)*[7]

Quatre groupes de documents ont été identifiés et constituent les quatre volets du projet aux contenus très différents : les thèses numériques, les publications commerciales numériques, les publications officielles numériques de la Confédération et les sites web d'importance patrimoniale. C'est ce dernier volet, nommé Archives Web Suisse, qui nous préoccupera dans ce chapitre.

Le système informatique pour l'archivage des documents est largement inspiré du modèle de référence OAIS (Open Archival Information System)[8]. Les différentes structures de ce modèle sont modulables et donc calibrables pour l'environnement des collectes de la BN. (BN 2012b)

Enfin, depuis 2012, le projet e-Helvetica est intégré comme service au sein même de la BN. Le service œuvre pour les traditionnelles missions d'un service d'archive : la constitution des collections, le catalogage, l'archivage à long terme et la mise à disposition des documents. C'est précisément ces deux dernières missions qui cristallisent les défis futurs auxquels est soumise la BN. (BN 2012b) En effet, on imagine aisément la nécessité pour les bibliothécaires et les partenaires associés de devoir réfléchir aux questions cruciales liées à la pérennisation des collections établies et à leurs modes d'accès. Certaines de ces questions seront notamment abordées au sous-chapitre 6.3.1.

4.1.1 Cadre légal

Les tâches et l'organisation de la BN sont réglées juridiquement par la LBNS. Cette loi qui encadre notamment les mandats de l'institution déclare à l'article 3, al. 1 que « La Bibliothèque nationale collectionne les informations imprimées ou conservées sur d'autres supports que le papier[9] […] » (Suisse 1992) : cette disposition suffisamment

[7] L'ordonnance sur la Bibliothèque nationale suisse du 14 janvier 1998 précise, notamment, les contours du mandat de collection de la BN concernant les Helvetica, à l'article 2 : https://www.admin.ch/opc/fr/classified-compilation/19980041/index.html

[8] Pour davantage de précisions (notamment le texte de la norme) à propos du modèle OAIS largement répandu aujourd'hui dans les services d'archives, consulter : http://www.archivesdefrance.culture.gouv.fr/gerer/archives-electroniques/standard/norme-oais-iso-14721/

[9] C'est moi qui souligne.

générale et abstraite inclut désormais les publications nées numériques, comme les e-books, les e-journals et les sites web (BN 2012b).

A l'inverse de la situation française, la Suisse ne dispose pas d'un arsenal juridique instituant un dépôt légal au niveau national. Les cantons sont responsables de légiférer à leur niveau, s'ils le souhaitent : c'est le cas de Vaud, Genève et Fribourg. Afin de mener à bien les missions qui lui sont néanmoins dévolues par la LBNS, la BN a signé une convention avec deux associations d'éditeurs suisses : l'Association Suisse des Diffuseurs, Editeurs et Libraires (ASDEL) et la Schweizer Buchhändler- und Verleger-Verband (SBVV). Cette convention stipule que les éditeurs membres de ces deux associations sont tenus de déposer un exemplaire de leurs publications auprès de la BN. (BN 2011)

Cette absence de dépôt légal va structurer les logiques archivistiques à l'œuvre au sein du volet Archives Web Suisse. Un cadre juridique comme celui du dépôt légal du numérique en France demeure particulièrement facilitateur, comme le souligne Barbara Signori, responsable e-Helvetica à la BN : « L'archivage web se trouve simplifié en termes d'obtention des droits. Lorsque la demande de collecte tombe, on économise des ressources. » (2015a). Cet état de fait impose à la BN de solliciter systématiquement chaque producteur de sites web sélectionnés par les bibliothécaires en charge de l'identification des contenus. Un courriel explicitant les objectifs d'Archives Web Suisse ainsi que le processus de collecte (« harvesting ») est envoyé à l'exploitant qui peut dès lors refuser le moissonnage de son site. Ce genre de scénario peut arriver, « […] mais à un très petit pourcentage. » (Signori 2015a). Enfin, la possibilité d'annoncer son site au service de coordination d'Archives Web Suisse reste une possibilité pour les éditeurs.

4.1.2 Archives Web Suisse

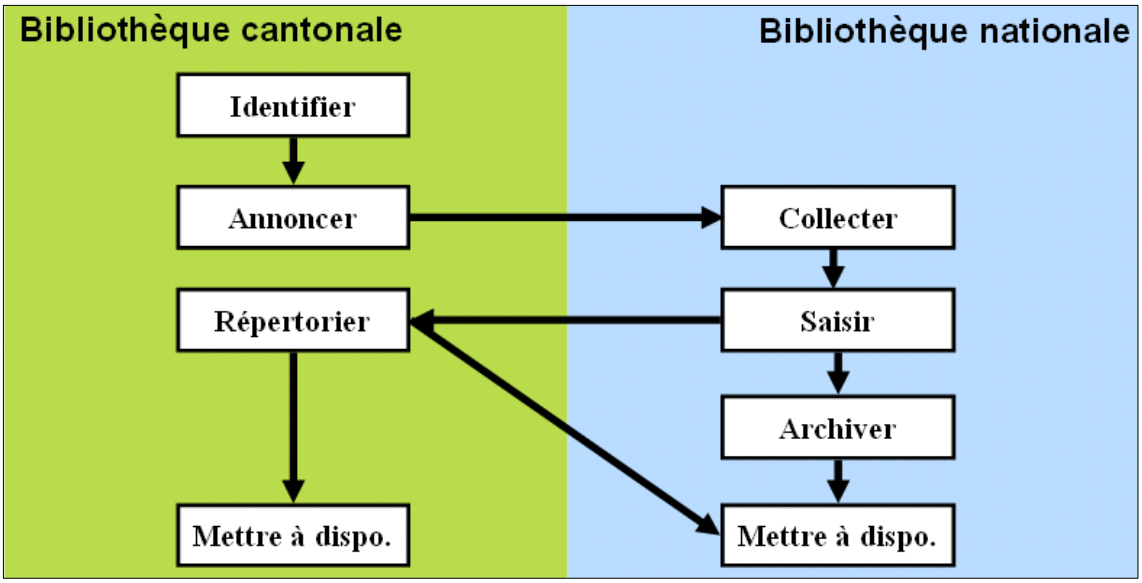
Le volet baptisé Archives Web Suisse a pour but de collecter les sites Internet patrimoniaux et non commerciaux (Balzardi 2008) de la Suisse. Il s'agit, dans le prolongement des collections déjà établies par la BN, de conserver et mettre à disposition le « patrimoine intellectuel de la Suisse » (BN 2012a). Le partenariat avec certaines bibliothèques cantonales/spécialisées dans le travail de sélection des sites web à valeur patrimoniale est au cœur de ce programme ; cette collaboration est basée sur l'article 10 (Coopération et coordination) de la LBNS. Ainsi, ce sont ces institutions qui sont chargées d'identifier et de sélectionner les sites web qu'elles font remonter auprès de la BN, au moyen d'un formulaire d'annonce en ligne. Cette étape de sélection et de pré-saisie sera explicitée au sous-chapitre 4.1.5, au travers de

l'exemple de la bibliothèque cantonale et universitaire vaudoises (BCU). La BN, quant à elle, est ensuite chargée de la collecte, du catalogage dans Helveticat, de l'archivage et de la mise à disposition.

Comme nous le verrons plus loin, l'approche archivistique de la BN concernant les documents numériques demeure fondamentalement « sélective et thématique » (Chaimbault 2008). Le cadre légal ne permettant pas l'exhaustivité d'un nom de domaine national, la BN a mis en place des processus de travail qui placent au centre les bibliothécaires et leur expertise documentaire.

4.1.3 Processus de travail

Schéma des responsabilités des tâches de la BN



(Signori 2015b)

Comme évoqué plus haut, ce sont certaines bibliothèques cantonales/spécialisées (une trentaine d'établissements) qui se chargent d'identifier et d'annoncer les sites sélectionnés. Il s'agit des deux tâches essentielles dévolues aux bibliothèques partenaires. Ainsi, elles commencent par identifier et sélectionner les sites web patrimoniaux de leur canton respectif selon des directives de collectes[10] qui ont été définies en commun avec l'ensemble des bibliothèques partenaires. Barbara Signori rappelle que ces principes de collecte sont « si nécessaire, […] aussi révisés en commun. » (2015a). Au moyen d'un formulaire en ligne, les bibliothèques annoncent dans un second temps les sites identifiés à la BN. Ce formulaire contient toute une

[10] L'ensemble de ces grands principes de sélection et d'exclusion est disponible au sein du document consultable ici : https://www.nb.admin.ch/nb_professionnel/01693/01699/01873/01895/index.html?lang=fr

série de métadonnées sur le site à archiver qui participent déjà à l'enrichissement de la future notice de la ressource.

La BN, désormais en possession des listes de sites web annoncés, s'adresse aux producteurs de ces sites pour obtenir leur autorisation et les droits associés. L'étape cruciale de la collecte à proprement parler (ou « harvesting ») peut ainsi commencer. Le principe technique de la collecte demeure assez simple et similaire à beaucoup d'autres programmes d'archivage du web : « […] depuis une page de départ, tous les liens sont suivis et les fichiers présents dans le domaine de collecte défini sont téléchargés. » (Locher 2015). Les pages privées ou protégées par un mot de passe ne peuvent pas être collectées. Cette étape est prise en charge par le logiciel open source Webspider Heritrix[11]. Les métadonnées qui ont été transmises par les bibliothécaires au moyen du formulaire sont saisies automatiquement et directement chargées sous forme de notices dans Helveticat. Si les bibliothèques cantonales partenaires le souhaitent, elles peuvent réclamer les notices à la BN afin de les intégrer à leur propre catalogue. (Signori 2015b). Les sites web qui ont été moissonnés par la BN sont ensuite stockés et archivés au sein d'un système de mémoire à long terme nommé Ninive. Chaque site possède un identificateur unique (URN) pour qu'il soit retrouvable au sein des deux composantes du système Ninive.

Enfin, l'ultime étape du processus réside dans la mise à disposition de la collection Archives Web Suisse : elle est accessible depuis les locaux de la BN et depuis peu, Barbara Signori indique que: « L'accès est aussi possible dans les locaux des bibliothèques partenaires, pour autant qu'elles aient effectué les installations nécessaires. » (Signori 2015a)

4.1.4 Périmètre et modes de la collecte

Dans la tradition des Helvetica, la collection Archives Web Suisse regroupe en grande majorité des sites web patrimoniaux de la Suisse. La question posée en filigrane est bien celle de la valeur patrimoniale d'un site internet : quelles réalités recouvre la notion de patrimoine ? Comme interpréter la valeur d'un site ? Les contenus numériques, de par leur nature, questionnent-ils le périmètre des Helvetica ? Comment identifier les (bons) sites ? Comment identifier les jugements normatifs qui peuvent présider à la sélection de tel ou tel site ?

Comme évoqué plus haut, le périmètre de la collecte prise en charge par la BN a été, dans une tradition fédéraliste, décidé collégialement avec des représentants des

[11] Il s'agit du même logiciel utilisé par la BnF et développé au sein du Consortium IIPC. Nous verrons plus loin dans ce travail la place qu'occupe cet organisme particulier.

différentes institutions partenaires. S'inscrivant dans les mandats de collection prévus par la LBNS et ne disposant pas d'un dépôt légal national, la BN a opté pour une approche « sélective et thématique », selon la terminologie adoptée par Thomas Chaimbault, rejetant tout projet d'exhaustivité. Il s'agit donc « […] d'archiver que des portions définies du web […] selon des critères spécifiques. […] cette approche vise à prendre des instantanés des sites à des intervalles précis. » (Chaimbault 2008). La dimension thématique quant à elle renvoie à « […] la constitution d'une collection de site web à l'occasion d'un événement particulier. » (Chaimbault 2008). Ce sont donc ces deux approches conjuguées qui forment les grands modes de sélection de la BN.

Ces deux modes de collecte sont également complétés par des grands principes de sélection (et d'exclusion) réunis au sein d'un document de travail (Signori 2011). Cette base de critères de sélection pose un premier cadre qui peut être complété et détaillé par chacun des établissements partenaires. Ce document exclut d'entrée de jeu certains sites pour des raisons techniques : la quantité de données pouvant être récoltées est limitée, l'utilisation de Flash et de JavaScript sur certains sites peut empêcher la collecte de l'intégralité de leurs contenus, des robots .txt empêchent parfois l'accès à l'archivage complet du site. Enfin, les crawler traps sont des applications générant à l'infini de nouvelles URL, de sorte que le crawler les suit également à l'infini et ne peut archiver le « harvesting ». (Signori 2011). Une longue liste de documents variés est également exclue de la collecte ; on y trouve, pêle-mêle, les jeux, les sites pornographiques, ainsi que tous les sites/pages qui contiennent des informations ou échanges privés.

En revanche, les sites web qui répondent à la définition d'un Helvetica comme vu plus haut rentrent dans le champ de la collecte. Les critères de sélection sont également attachés à la constitution de certaines collections, notamment celle concernant les cantons. Il s'agit de constituer une collection de site web représentant le canton en tant que collectivité territoriale ; on y trouve notamment ses propres sites officiels, ou tout autres sites d'importance patrimoniale pouvant témoigner d'une dimension culturelle particulière dudit canton. Parallèlement, la BN sélectionne des sites web sur des événements spéciaux en Suisse, constituant alors une collection en soi. C'est le cas, par exemple, d'une collection sur les élections de 2007. (Signori 2011)

D'autres indices, s'ils forment un faisceau, peuvent constituer des critères de sélection déterminants. Suivant s'il s'agit de sites web représentants (autour d'une personne) ou thématiques (autour d'une collectivité), ces critères varient mais renvoient toujours à la qualité et la densité du contenu, au sérieux de son producteur ou encore au

professionnalisme de sa présentation. La BN propose ainsi aux sélectionneurs des institutions cantonales/spécialisées une grille d'évaluation leur permettant d'encadrer leur sélection. Les trois axes de cette grille sont comme une synthèse des principes de sélection à l'œuvre : contenu, navigation, structure et conception. La pondération élevée associée au volet « contenu » de la grille souligne son importance.

L'une des particularités fondamentales du programme Archives Web Suisse est la profondeur de l'archivage pratiquée. En effet, la BN propose un archivage plutôt profond du site en intégrant « […] tous les contenus d'un site, ainsi que les commentaires, ceci pour autant que cela soit techniquement possible et que la grandeur maximale pour l'archivage ne soit pas dépassée. » (Signori 2015a). Le but ultime étant de pouvoir naviguer à l'intérieur du site comme il se présentait à un moment t. De plus, c'est également le contexte du site qui est sauvegardé au travers de l'archivage des liens sortant du site qui disent ainsi quelque chose d'un écosystème plus global. Cette exigence de qualité et de profondeur est nécessaire, au risque de perdre le site et/ou de ne pas l'archiver. La profondeur de l'archivage est parfois contrecarrée par des limites techniques, comme évoqué précédemment. Les contenus dynamiques omniprésents aujourd'hui sur le web bloquent parfois les robots dans leur travail de collecte. Ces freins techniques éventuels compromettent le travail d'archivage :

> *« Lorsque nous ne pouvons pas collecter un site pour des raisons techniques, ou lorsque nous ne pouvons pas le collecter de manière suffisante, nous ne l'archivons pas. Lorsque nous ne pouvons pas afficher un site que nous avons cependant réussi à collecter complètement, nous l'archivons. » (Signori 2015a)*

Barbara Signori souligne les qualités et les forces de l'approche archivistique de la BN en déclarant : « La sélection garantit une certaine qualité des sites et nous avons connaissance de ce que nous avons enregistré dans les Archives Web. » (2015a). Il est certain que la profondeur de l'archivage assurée et la politique documentaire établie en amont de la sélection assurent un bon niveau de qualité ? des collections. Si les robots n'interviennent pas dans la sélection, la mise en œuvre de la sélection par les différents responsables cantonaux n'est pas sans poser plusieurs questions que nous avons soumises à la responsable de l'archivage du web pour le canton de Vaud.

4.1.5 Responsabilité des bibliothèques cantonales : le cas vaudois[12]

Pour mieux saisir les logiques de travail autour du processus concret de sélection des sites et les enjeux à la fois patrimoniaux et professionnels qu'il pose, nous avons rencontré Brigitte Steudler, responsable de la Documentation vaudoise au sein de la Bibliothèque cantonale et universitaire de Lausanne (BCU) et personne de contact opérationnel dans le cadre d'Archives Web Suisse pour le canton de Vaud. Ce canton est aujourd'hui le deuxième plus représenté avec plus de 450 sites archivés pérennes depuis 2010. Outre son propre travail de sélection, Brigitte Steudler centralise toutes les propositions qui lui sont soumises par différents partenaires du canton ou par certaines institutions, avant de les faire remonter à la BN. Elle constitue ainsi un réservoir à annoncer avec environ septante sites chaque année, mais ceux-ci sont archivés petit à petit avec un inévitable décalage dans le temps, en raison de la lourdeur du « harvesting ». Chaque site qui est remonté auprès de la BN est accompagné d'un bordereau descriptif qui liste une série de métadonnées, utiles dans la perspective du futur catalogage de la ressource sur Helveticat. Il peut arriver que la BN renvoie une note à l'institution collecteuse, stipulant que tel site est trop volumineux pour une collecte ou que tel autre rentre en conflit avec les directives de collecte évoquées plus haut ; une discussion peut alors être engagée au cas par cas. L'un des points de discussion apparu au fil du temps est la question de la territorialité du site à archiver : un site documentant le canton de Vaud peut être basé à Genève. Malgré le caractère transterritorial (ou transnational) du web, le cas de ces site a dû être discuté : quel canton doit le prendre en charge ? D'autres discussions peuvent également survenir lorsque la BN, ponctuellement, impose une collecte ciblée autour d'un événement particulier, comme le « XIIIème Sommet de la Francophonie » de Montreux en 2010 : ce type d'événement documente-t-il réellement le canton ? Par ailleurs, pour éviter une trop grande disparité entre les différentes collectes cantonales, la BN peut freiner les soumissions de sites des bibliothèques en les enjoignant de payer leur archivage.

Déchargée de toutes les questions et problèmes techniques qui sont du ressort de la BN, l'intervention de Brigitte Steudler est donc purement qualitative, documentaire et intellectuelle. En effet, comme évoqué au sein du chapitre précédent, son mandat est de documenter le canton de Vaud à toutes les époques, dans tous les domaines, sur

[12] L'entier de ce chapitre est une synthèse de l'entretien avec Mme Brigitte Steudler, responsable de la Documentation vaudoise, personne de contact opérationnel d'Archives Web Suisse pour Vaud au sein de la Bibliothèque cantonale et universitaire de Lausanne (BCU), Lausanne, 16 avril 2015.

tous les supports. Ce périmètre extrêmement large offre une marge interprétative dans la collecte des sites signifiants : il est donc possible de trouver des échappées hors du mandat purement institutionnel et territorial du projet. En filigrane du processus de sélection, la question de la valeur patrimoniale est en jeu : comment mesurer le potentiel mémoriel et patrimonial d'un site ? Quels types d'informations doit-il contenir ? Quelles sont les sources qui documentent, racontent, thématisent le canton de Vaud ? Si ces questions sont sans cesse discutées avec ses différents partenaires, Brigitte Steudler opère des choix, jalonnés par quelques principes essentiels dans son appréhension de la masse du web, discutés lors de notre entretien : la nécessité du contenu, la création artistique, les initiatives individuelles et enfin, la notion de « document en danger ».

L'un des enjeux fondamentaux est la présence de contenus importants ou d'archives au sein du site : il faut que le site, même s'il concerne une institution culturelle, regroupe des contenus ou des archives. L'inconnue de la date à laquelle le site sera effectivement collecté par le robot peut poser problème : les documents repérés en amont y figureront-ils toujours ? Malheureusement, ce problème temporel ne peut être réglé par une accélération des procédures concernant certains sites. Par ailleurs, face à certains sites extrêmement profonds et/ou possédant des archives particulièrement imposantes (comme le site de l'EPFL), c'est le producteur du site qui détermine les parties de celui-ci à archiver.

La création artistique est, selon Brigitte Steudler, un axe cardinal de la documentation du patrimoine : ces documents représentent fondamentalement le canton et son histoire. Dans cette perspective, le site d'un photographe ou d'un dessinateur qui n'est pas publié serait une source intéressante. Elle souligne également l'intérêt des sites de particuliers et d'amateurs qu'elle oppose à la pauvreté de certains sites institutionnels, volontiers collectés par principe alors qu'ils ne constituent parfois que de simples annuaires. A titre d'exemple, le site personnel d'un passionné de la CGN ou d'associations diverses peut constituer un précieux témoignage de la vie du canton. Ainsi, Brigitte Steudler appuie ici les propos d'un chercheur interrogé dans le cadre d'une étude menée par la BnF sur les attentes des utilisateurs des archives du web qui souligne également l'importance des ressources des particuliers : « […] le savoir contenu […] dans les « *pratiques amateurs* » : celles des amateurs et des passionnés, proches de la retraite, « *les gens qui ont une mémoire de ça* », et prennent le temps de « *remettre leurs souvenirs en ordre* ». » (Chevallier, Illien, 2011, p. 19)

Une troisième dimension intervient dans le processus de collecte de Brigitte Steudler : celle de « document en danger ». L'urgence de préserver des sites importants qui risquent de disparaître constitue un défi majeur. Ainsi, le critère de la volatilité, du document dont on risque de ne pas pouvoir garder la trace lui semble fondamental, malgré une certaine lourdeur technique, notamment concernant certains contenus audiovisuels. Face à l'urgence de se saisir de certains documents, Brigitte Steudler s'étonne notamment du rejet pur et simple des blogs au sein des directives de collecte : le blog est selon elle le journal manuscrit du XVIIIe, de l'écrivain qui au jour le jour écrit ses pensées. Néanmoins, cet état de fait semble sur le point de changer selon Barbara Signori : « Les blogs seront autorisés prochainement » (2015a). Dans le projet de pouvoir documenter le plus fidèlement possible les générations futures, Brigitte Steudler est soucieuse de pouvoir conserver tout ce qui n'est pas édité, à l'instar des artistes vaudois présents sur la plateforme Mx3[13]. Ce type d'initiatives aboutit parfois difficilement, se heurtant aux jugements de valeur documentaires de certains collaborateurs de l'institution. Car s'il s'agit pour Brigitte Steudler de mettre un point d'orgue à ne jamais se poser en arbitre ou en juge face à la variété des ressources du web (et à une éventuelle « valeur intrinsèque »), il demeure difficile de résister face à la violence symbolique exercée par l'institution et aux jugements de valeur qu'elle émet.

Les résistances face au projet d'Archives Web Suisse se logent parfois au sein même de l'institution : ainsi, à titre d'exemple, c'est dans la difficulté à déployer des infrastructures d'accès pour les nouvelles collections du web ou encore dans le peu de communication pour valoriser ces nouvelles ressources que l'on peut percevoir certains freins à l'essor du projet. Les craintes liées aux efforts techniques que ces nouveaux contenus impliquent, expliquent peut-être certaines attitudes timides. Par ailleurs, les questions et les résistances ne se cristallisent pas uniquement au sein de l'institution, mais également auprès d'un certain public qui ne comprend pas pourquoi des ressources issues du web sont accessibles seulement depuis les locaux des institutions. Il s'agit d'un frein majeur à la consultation de ces archives nouvelles qui cherchent encore leur(s) public(s) et leurs modes d'accès.

[13] Il s'agit de "la plate-forme musicale de DRS3, Virus, Couleur3, Rete3 et Radio Rumantsch. Les musiciens peuvent y présenter leurs morceaux aux cinq rédactions musicales et au reste du monde. Les fans, organisateurs et labels peuvent s'y inscrire et partir à la découverte de la création musicale suisse."
MX3, 2014. *Mx3* [en ligne]. 2006-2014. [Consulté le 2 mai 2015]. Disponible à l'adresse: www.mx3.ch

4.2 BnF : Archives de l'internet

Afin de saisir les contours de l'entreprise d'archivage du web entreprise par la BnF, nous allons, dans un premier temps, nous concentrer sur son cadre légal particulier qui lui a permis de se déployer et tenter de saisir ses rouages. Nous évoquerons ensuite rapidement les outils techniques et certains de leurs usages. Enfin, nous nous concentrerons sur les périmètres et modes de collectes variés qui constituent autant d'outils pour appréhender la variété documentaire née numérique. Les logiques de travail et les conceptions professionnelles seront notamment abordées.

Il est à noter que nous nous concentrerons spécifiquement sur le mandat de la BnF et évacuerons celui de l'Institut national de l'audiovisuel (INA) chargé de collecter et de stocker spécifiquement des

> *« sites de médias audiovisuels, des sites qui enrichissent ou documentent les contenus de ces médias – comme les sites officiels de programmes mais aussi les blogs ou sites de fans,* [ainsi que] *des sites des services de médias audiovisuels à la demande […] »* *(Mussou 2012)*

Les logiques de travail de l'INA sont très spécifiques aux médias collectés et exigeraient un travail à part entière sur cet organisme mandataire. Par ailleurs, l'ancienneté de la BnF et l'étendue de son mandat dans le cadre des Archives de l'Internet en font un objet d'étude potentiellement plus riche, soulignant davantage les différents défis de l'archivage d'un web national aujourd'hui.

4.2.1 Le Dépôt légal du numérique : un cadre légal

Depuis son instauration en France par l'ordonnance de Montpellier en 1537, le dépôt légal français n'a cessé de s'adapter successivement à tous les supports informationnels reflétant la mémoire de la production éditoriale et culturelle française. Dans cette logique, la BnF s'est dotée, depuis le mois d'août 2006, d'un cadre juridique qui étend cette fois-ci le pluriséculaire dépôt légal français[14] aux publications de l'Internet : il s'agit du dépôt légal numérique. Cette disposition récente, qui a engagé presque sept ans de travail, repose sur un « support législatif » de la Loi relative au droit d'auteur et aux droits voisins dans la société de l'information (Dadvsi)[15]. Le dépôt légal du numérique a ainsi intégré les dispositions du Code du patrimoine (articles L

[14] Ce mécanisme « édicte que toute publication produite ou diffusée en France doit entrer dans les collections nationales. » (Bonnel, Oury, 2014, p. 2) Pour plus de détails concernant le dépôt légal français dans son acception générale, consulter: http://www.bnf.fr/fr/professionnels/depot_legal_definition/s.depot_legal_mission.html

[15] La Dadvsi no 2006-961 du 1er août 2006 est consultable dans son intégralité sur le site : http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000266350

131-1 et suivants)[16]. La loi définit l'élargissement du dépôt légal au numérique en ces termes :

> *« Sont également soumis au dépôt légal les signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique[17]. »* *(France 2015)*

Désormais, le dépôt légal qui concernait essentiellement le papier s'étend aux publications numériques. Selon Clément Oury, chef du service du Dépôt légal du web à la BnF, la définition « générale et abstraite » que propose la loi l'est à dessein; elle lui permet de s'adapter au futur des technologies et de combler ainsi le temps de latence entre les innovations technologiques et le temps du législateur. (2012)

Le décret d'application de cette loi est quant à lui publié le 19 décembre 2011[18]. Ce dernier précise plus spécifiquement l'objet du dépôt qui reste très large dans le texte de loi : « Il définit […] ce que l'on doit entendre comme l'internet français : […] des sites hébergés sur des « domaines de haut niveau » français (*.fr*, *.paris*, *.re*, etc.) ; et/ou des sites dans un nom de domaine enregistré par une personne domiciliée en France ; et/ou enfin des sites produits sur le territoire français. » (Bonnel, Oury, 2014, p. 3). De plus, le décret investit deux institutions de la mission de conserver l'Internet français : l'INA pour les sites de télévision et de radio ou « principalement consacré » et la BnF pour tout le reste du web. Enfin, le texte évoque quelques modalités de collecte : « […] tous les noms de domaines doivent faire l'objet d'une collecte ; […] la profondeur de collecte n'est pas précisée et l'exhaustivité de la collecte de chaque site n'est pas demandée. […] une fréquence minimale d'archivage [d'] une fois par an […]. » (Bonnel, Oury, 2014, p. 3). Nous développerons plus loin l'influence de ce décret sur les logiques, les modèles et les périmètres de collecte. Afin de mener à bien les missions qui incombent désormais à ce nouveau champ de collecte, un service du dépôt légal numérique est désormais en place au sein du Département du dépôt légal de la BnF.

L'une des caractéristiques fondamentales du dispositif du dépôt légal français est son caractère non-sélectif et encyclopédique, comme le précisent Sylvie Bonnel et Clément Oury : « […] toute production culturelle a vocation à être déposée, quelle que soit la « valeur » que les bibliothécaires lui attribuent. ». (2014, p. 2). Le dépôt légal, qu'il soit

[16] Les dispositions du code du patrimoine sont consultables dans leur intégralité sur le site : http://www.legifrance.gouv.fr/affichCode.do;jsessionid=77BCE731A86D0A1C02D19877FC37F3C7.tpdila20v_2?idSectionTA=LEGISCTA000006159934&cidTexte=LEGITEXT000006074236&dateTexte=20150429

[17] C'est moi qui souligne.

[18] Le décret d'application est consultable dans son intégralité sur le site : http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000025002022&dateTexte=&categorieLien=id

numérique ou non, fonctionne ainsi comme un miroir de la société française et de ses productions, indépendamment de la qualité de ses contenus :

« […] the philosophy of legal deposit is indeed to keep a record of the « best » along with the « worst » as collections should be a mirror of society's global cultural production and evolution over centuries. » (Lupovici et al. 2006, p. 2)

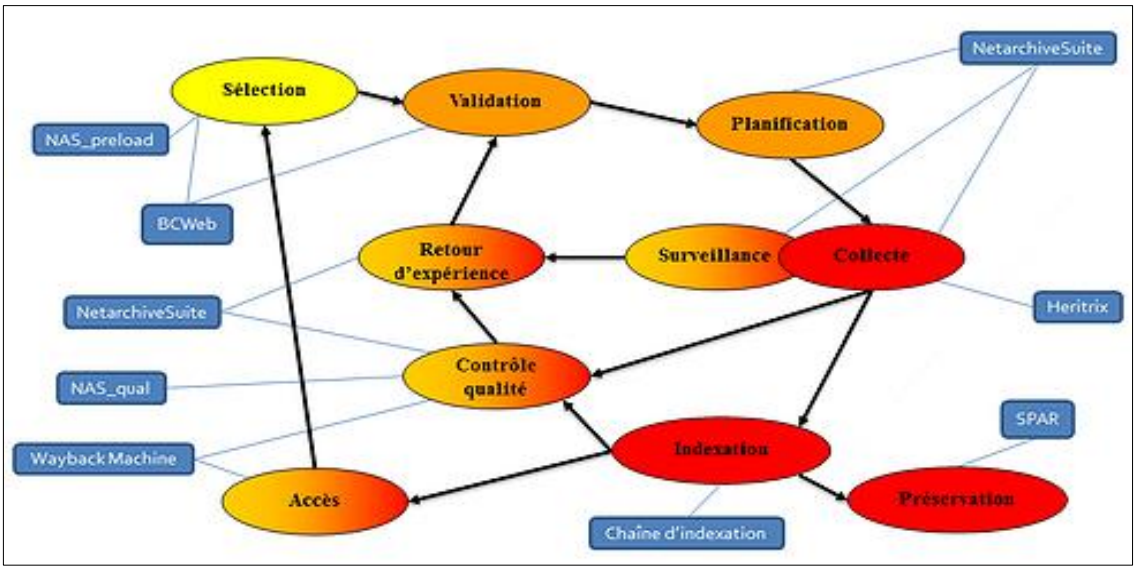
Contrairement au contexte helvète, il est à noter que le texte de loi autorise les institutions mandataires à moissonner des contenus sans le consentement de leur éditeur ; il aurait été en effet impossible d'exiger des éditeurs numériques de systématiser un dépôt à chaque création de contenu. Cette particularité constitue une « exception au droit d'auteur et aux droits voisins » (Illien 2008) qui est contrebalancée par des conditions d'accès aux collections très restrictives qui protègent ainsi les ayant-droit. Nous observerons plus loin dans ce travail comment ces restrictions empêchent les archives de rencontrer leurs publics.

Appliquer la mécanique du dépôt légal à la variété des productions et contenus de la toile n'est pas sans receler de nouveaux défis et poser certaines interrogations : comment conserver le plus largement possible une production aussi vaste et volatile que celle engendrée par Internet ? Ou encore, comment conjuguer les exigences techniques et budgétaires qu'impliquent les programmes d'archivage du web? Des pratiques documentaires, un périmètre de collecte, des leviers techniques et des politiques documentaires vont tour à tour répondre aux injonctions du dépôt légal du numérique.

4.2.2 Pratiques et outils technologiques

L'entreprise d'archivage du web français nécessite une adaptation des outils de travail du cycle complet d'archivage aux réalités du web. Si le flux de traitement des documents reste le même, ce sont bien le tempo des tâches et les outils qui diffèrent. (BnF 2015a) La BnF propose un schéma global des tâches et outils nécessaires au bon fonctionnement de l'archivage du web :

Schéma général des flux de production de l'archivage du web à la BnF



(BnF 2015b)

Les grandes étapes du flux de production que sont la sélection, la validation, la planification, la collecte, l'indexation, la préservation et l'accès sont encadrées par une batterie d'applications que nous détaillons ici.

L'application BnF Collecte du web (Bcweb) permet aux bibliothécaires et autres agents partenaires sélectionnant les sites web de gérer des listes de sites à collecter. Ces listes sont ensuite périodiquement transmises aux robots afin qu'ils les moissonnent. Les sites sélectionnés pour leur intérêt thématique par le biais de cette application s'inscrivent dans le cadre de collectes ciblées impliquant un archivage profond.

Dans le cadre des collectes larges, l'application NAS_preload permet de rassembler et de préparer l'ensemble des listes des noms de domaines collectés (*.fr*, *.re*, *.nc*) par le biais, notamment, de l'Association Française pour le Nommage Internet en Coopération (AFNIC) et l'OPT-NC. L'application, une fois ces listes versées, va procéder à une série de tests pour déterminer d'une part si les sites existent réellement, et d'autre part les codes http renvoyés. Ainsi, seuls les sites actifs seront moissonnés par les robots. Par ailleurs, l'application constitue également un outil statistique de l'activité des sites du domaine français. (BnF 2015a)

L'outil NetarchiveSuite peut être compris comme une application permettant la planification, le suivi et le contrôle qualité des collectes et des archives constituées. Il peut être utilisé autant face à de petits volumes de données qu'à des très grands ensembles de documents. (BnF 2015a)

Heritrix est le robot de collecte utilisé par la BnF. Il s'agit d'un logiciel extrêmement souple qui peut fonctionner autant comme un module isolé ou alors se rattacher à un outil comme NetarchiveSuite. Le robot « propose plusieurs types de filtres, d'extracteurs et de processus modulaires selon les besoins de collectes. » (BnF 2015a). Des ajouts de scripts sont également possibles. Le robot stocke les données sur des fichiers ARC ou WARC.

NAS_quai propose un panorama d'indicateurs de production, destinés aux équipes du dépôt légal du numérique. On y trouve, par exemple, le poids des données, le nombre d'URL collectées, le code réponse http, etc. Tous les chiffres issus des indicateurs sont analysés puis mis à disposition au sein de l'Observatoire du dépôt légal. (BnF 2015a)

Lorsque les fichiers ont été déposés par le robot Heritrix au sein des fichiers containeurs ARC ou WARC, des scripts internes permettent un processus d'indexation des sites. Cette étape est particulièrement essentielle dans la perspective de la mise à disposition des collections. (BnF 2015a)

Relativement connue du grand public car initialement conçue par Internet Archive, l'application Wayback Machine est l'interface d'accès privilégiée de nombreuses bibliothèques. Le logiciel permet notamment de naviguer dans les collections « comme à l'époque où [elles] ont été collectées » (BnF 2015a), ainsi que de comparer différentes versions des sites (ou des pages) dans le temps. (BnF 2015a)

Depuis 2013, les archives sont versées au sein d'un entrepôt sécurisé nommé SPAR. C'est dans cet espace que l'intégrité des fichiers est garantie. L'une des utilités de cet entrepôt est la surveillance de l'évolution des formats des fichiers préservés, permettant de lutter contre les éventuels problèmes d'obsolescence. Il s'agit ensuite, suivant les cas, de procéder ou à une migration ou à une émulation. (BnF 2015a)

Il est à noter enfin que beaucoup de ces applications et outils sont des logiciels libres, comme le soulignent Sylvie Bonnel et Clément Oury : « La plupart des outils de collecte et d'accès aux archives du web utilisés par la BnF sont des outils *open source* […]. » (2014, p. 4). Ces outils sont le fruit de la coopération internationale en matière d'archivage du web : « [les outils] ont été développés au sein du consortium international pour la préservation de l'internet, ou IIPC. » (Bonnel, Oury, 2014, p. 4). Cette batterie de ressources vient souligner en filigrane l'impérieuse nécessité d'une coopération internationale généralisée et les résultats, ici technologiques, qu'elle engendre.

4.2.3 Périmètre et mode des collectes

Le décret d'application de décembre 2011 circonscrit les pratiques de la BnF et précise le périmètre d'une collecte a priori vertigineuse, dans une approche à la fois exhaustive, sélective et thématique. En tout premier lieu, nous pouvons évoquer le périmètre associé aux noms de domaine. Comme spécifié plus haut, les « contenus édités ou les éditeurs doivent avoir un lien avec le territoire français » (BnF 2014) ; cela implique bien entendu l'ensemble des sites *.fr* (et ceux des territoires d'outre-mer), mais également d'autres sites enregistrés en *.com, .org, .net*, etc. par des éditeurs qui sont domiciliés sur le territoire français. Certains contenus produits par des personnes morales étrangères sur le territoire français sont également pris en compte. (BnF 2014) Concrètement, la BnF collabore directement avec l'AFNIC qui lui fournit l'ensemble des sites en *.fr* : environ 2,8 millions[19] de noms de domaine[20]. Il est également défini dans la loi qu'il s'agit de contenus « […] faisant l'objet d'une communication au public […]. » : cette précision exclut ainsi d'entrée de jeu toutes les communications privées. Ainsi, les courriels, les réseaux sociaux privés[21] ou encore les forums de discussions restreints ne sont pas soumis à la collecte.

4.2.3.1 La collecte large

Comme évoqué précédemment, la question de la profondeur de la collecte du site est au centre des enjeux liés à son périmètre. Aucun degré de profondeur ou d'exhaustivité de collecte n'est exigé par le décret d'application de la loi. Pour des raisons évidentes de stockage et techniques, il serait impossible de moissonner le web français dans toute sa profondeur. Le premier modèle de collecte de la BnF, la « Collecte large », moissonne ainsi le web dans son intégralité, mais superficiellement. C'est ici une rupture fondamentale avec la tradition du dépôt légal qui vise l'exhaustivité de la production éditoriale : nous passons d'une logique de l'exhaustivité à celle de la représentativité (Oury 2012), en constituant « une image incomplète, mais fidèle, de l'internet français […]» (Bonnel, Oury, 2014, p. 4). Néanmoins, Gildas Illien, tout en intégrant ce déplacement vers un souci de représentativité, nuance ce changement de paradigme : en effet, c'est bien cette notion de « fidélité » qui n'induit pas une rupture totale avec la tradition du dépôt légal : « Les collections constituées dans ce cadre répondent à la tradition française du dépôt légal, qui doit rester aveugle, au sens où il se veut plus représentatif que qualitatif. » (Illien 2008).

[19] Le détail des chiffres de l'AFNIC est disponible à l'adresse : https://www.afnic.fr, notamment dans l'espace « Ressources ».

[20] Il s'agit ici du nombre total de sites enregistré en *.fr* Le web français, quant à lui, est estimé dans son ensemble à sept ou huit millions de sites. (Bonnel 2014)

[21] Il est à noter que les parties publiques des réseaux sociaux, indexées par Google, peuvent être collectées par les logiciels-robots.

Cette collecte dite « large » est pratiquée une fois par an et capture plus de 90% des sites préalablement identifiés en partenariat avec l'AFNIC. Les logiciels robots (ou « robots moissonneurs ») parcourent le web de lien en lien sur tous les sites qu'ils rencontrent : ils fonctionnent « […] de la même manière que les robots indexeurs des moteurs de recherche […] » (Aubry 2008). Cette exploration du web se déploie soit en profondeur (liens entrant à l'intérieur d'un même site), soit en largeur (liens sortant vers d'autres sites) (Illien 2008). En termes volumétriques, cette collecte représente 50% des ressources prévues par la BnF, soit 50 téraoctets (To) (Bonnel, Oury, 2014, p. 4). Ces échantillons de surface représentaient environ quatre millions de sites lors de la dernière collecte.[22] Contrairement au tempo en vigueur au sein de la BN, le délai entre la capture d'un site et son indexation automatique est d'environ deux semaines ; celle-ci attribue une URL à chaque site et permet à l'usager d'accéder au site avec le logiciel « Wayback Machine ». Le nom exact de l'adresse est donc nécessaire pour accéder au site et pour ensuite y naviguer dans son environnement primaire, celui du « web vivant ». Aucune métadonnée n'est ainsi ajoutée aux sites lors de la collecte : cette réalité constitue une difficulté majeure pour l'accès des chercheurs à ces archives. L'indexation « full text » n'est mise en place aujourd'hui que pour moins de 5% des collections et, en raison des coûts importants qu'elle suppose, est davantage réservée aux collections impliquant des ressources humaines. Néanmoins, une partie du contexte du site est sauvegardée, puisque le robot collecte également, mais de façon très superficielle, les sites vers lesquels la ressource pointe : la page d'accueil d'un autre site, un texte ou une image, par exemple.

Comme le synthétise Gildas Illien, ce processus de collecte large constitue des :

« […] archives du web […] lacunaires puisqu'il peut manquer des fichiers, des pages, mais aussi parce qu'il est impossible de moissonner tous les sites en permanence : les collections constituées sont rarement des séries exhaustives ; elles se présentent plutôt comme des recueils de traces ou d'échantillons du web liés entre eux […]. »

(2008)

Cette collecte « large » impliquant immanquablement des lacunes et des collections fragmentées, d'autres types de collecte, que nous présentons aux sections suivantes, viennent les combler.

[22] Entretien téléphonique avec Mme Le Follic, Département du Dépôt légal numérique, Lausanne-Paris, 15 mai 2015.

4.2.3.2 Les collectes ciblées

« […] on ne peut s'en remettre entièrement à des robots pour constituer le patrimoine de demain. » (Illien 2008). L'avertissement de Gildas Illien est au cœur de la politique des collectes dites « ciblées ». Celles-ci, à l'inverse de la collecte large, visent à s'emparer de sites en profondeur et à une fréquence plus élevée. Cette sélection, produite par des bibliothécaires de la BnF (ou des partenaires externes) représente environ 30'000 sites, choisis pour leurs contenus particulièrement signifiants ou rattachés à une grande thématique : il s'agit des collectes dites « courantes ». Les sites web collectés peuvent également être rattachés à un événement particulier, comme ceux consacrés aux élections présidentielles de 2002, par exemple : il s'agit des « collectes projet ».[23] L'enjeu de ce volet de la collecte est double : il s'agit d'une part d' « […] apporter une valeur ajoutée à la collection constituée par robot […]» et d'autre part d' « […] assurer la valorisation des fonds auprès du public des chercheurs. » (Illien 2008).

Si les robots-logiciels sont ici exclus du processus de collecte, comment s'organisent les bibliothécaires et leurs différents partenaires ? La collaboration entre les différents agents est de mise dans le processus de sélection : une centaine de « correspondants DLweb », provenant des grands départements de collections thématiques de la BnF (ainsi que du Département du dépôt légal), se chargent ainsi de l'identification, du contrôle qualité et de la valorisation des sites web : ce sont eux qui fixent leurs propres critères de sélection et décident également de la fréquence et de la profondeur de la collecte pour chacun des sites identifiés. Bien entendu, la profondeur de collecte est liée à sa fréquence : un site particulièrement copieux ne peut être moissonné plusieurs fois par jour. Seuls certains sites de presse/média sont collectés jusqu'à trois fois par jour.[24] Les correspondants peuvent être secondés par des bibliothécaires ou des chercheurs partenaires. (Bonnel 2014) Une application pour les agents de la collecte a été mise en place : il s'agit de « BnF Collecte du Web » (ou « BCweb ») qui permet aux sélectionneurs de faire remonter leurs choix au moyen d'une fiche descriptive axée autour de la fréquence, de la profondeur et du budget adéquats.

Les collectes ciblées faisant intervenir d'importantes ressources humaines dans le processus de sélection, elles impliquent également, dans une tradition bibliothéconomique, la mise en place de politiques documentaires auxquelles les collaborateurs peuvent se référer. Ces dernières s'inscrivent autour des grands

[23] http://www.bnf.fr/fr/collections_et_services/livre_presse_medias/a.archives_internet.html
[24] Entretien téléphonique avec Mme Le Follic, Département du Dépôt légal numérique, Lausanne-Paris, 15 mai 2015.

principes fondateurs de la BnF, inscrits dans sa charte documentaire de 2005[25] : la France comme objet d'étude privilégié, l'encyclopédisme, la dimension temporelle avec comme point de mire des collectes s'inscrivant dans le temps. Ainsi, les collectes ciblées privilégient le domaine français *.fr*, font le pari de représenter la pluralité des grands champs de la connaissance et complètent les collections déjà existantes. (Bonnel, Oury, 2014, p. 7)

4.2.3.2.1 Les collectes courantes

Afin de mener à bien les collectes courantes, il s'agit de considérer le web et ses productions sous deux angles complémentaires : la continuité des collections et l'exploration de nouveaux territoires. En intégrant le web comme un prolongement documentaire des collections déjà existantes, la BnF conscientise la dimension patrimoniale d'Internet. Il ne s'agit pas pour elle de distinguer les contenus numériques natifs comme une source documentaire parallèle, mais bien de l'inscrire dans un continuum patrimonial en faisant des ressources du web un nouvel embranchement d'un arbre déjà planté en 1537. La production éditoriale, dans toute sa diversité, éclosant aujourd'hui majoritairement sur le web, il est nécessaire de l'intégrer dans les grandes diagonales des collections déjà établies par la BnF. Bonnel et Oury exemplifient cette réalité à travers le département des Cartes et Plans de la BnF qui considère et intègre les contenus de blogs d'amateurs de cartes dans ses collections. L'enjeu archivistique et patrimonial n'est évidemment pas celui de l'exhaustivité ou de la valeur intrinsèque du document, mais encore une fois celui de la représentativité, « […] de la manière dont on se représente l'espace à un moment donné, et des usages qui peuvent être retranscrits de manière cartographique. ». (Bonnel, Oury, 2014, p. 7)

L'exploration de nouveaux territoires incite les bibliothécaires à demeurer attentifs aux nouveaux types de contenus et d'organisations du savoir nés numériques et qui, pour cette raison, échappent a priori aux grandes classifications bibliothéconomiques. Les blogs et les forums constituent en ce sens des écosystèmes informationnels inédits et de nouveaux horizons documentaires très riches à conserver. Plus souterrains encore, les sites de Net Art ou plus largement représentant « […] des formes novatrices de création et/ou de diffusion sonore, audiovisuelle ou multimédia apparues avec l'Internet […] » (Bonnel, Oury, 2014, p. 7) sont également pris en compte par la BnF.

En termes de logique de travail, deux grandes approches complémentaires sont activées : la sélection et l'échantillonnage. La première d'entre elles mobilise des

[25] La charte documentaire de la BnF est consultable dans son intégralité à l'adresse : http://www.bnf.fr/documents/charte_doc_acquisitions.pdf

réflexes professionnels que l'on peut aisément associer à ceux de la bibliothéconomie : il s'agit de la sélection en amont de sites « […] habituellement sur la base d'un jugement de la qualité ou de la valeur scientifique ou esthétique du site. » (Bonnel, Oury, 2014, p. 6). Le choix du site s'opère comme celui d'un bibliothécaire feuilletant un catalogue d'ouvrages en vue de nouvelles acquisitions. (Bonnel, Oury, 2014, p. 6) C'est la logique appliquée au sein des différents départements acquéreurs de la bibliothèque. L'échantillonnage, en revanche, consiste à compiler le plus largement possible des sites, sans mesurer leur intérêt propre ou leur utilité future ; l'enjeu réside ici, dans une logique proche de celle du dépôt légal, dans le souci de représentativité d'une production globale par le biais d'échantillons. C'est la raison pour laquelle ce sont les différents départements gestionnaires du dépôt légal qui activent cette logique de travail particulière.

4.2.3.2.2 Les collectes projet

Les collectes projet se distinguent clairement des collectes courantes : en effet, elles visent à répondre à des attentes documentaires précises et s'inscrivent donc dans un périmètre plus restreint que les collectes courantes. (Bonnel, Oury, 2014, p. 7-8). Même si les thématiques des collectes projet semblent plus circonscrites que pour les collectes courantes, il n'en demeure pas moins que leur caractère transversal et leur lien fort avec l'actualité restent de mise. (BnF 2015b) Parfois réalisées en coopération avec des partenaires externes (bibliothèques, centres de recherches, associations), elles ne sont pas vouées à s'inscrire durablement dans le temps des collections. Gildas Illien pointe également un processus de communication et de « marketing » de ces collections :

> *« La constitution de ces corpus vise à réaliser des produits d'appel et des clés de valorisation au sein d'une collection si importante en volume qu'elle nécessitera des points d'entrée intelligibles et attractifs à destination des premiers archinautes. »* *(2008)*

La première collecte projet expérimentale mise en place par la BnF concerne l'élection présidentielle française de 2002 ; elle est exemplaire de toutes celles qui suivront, même si chacune d'entre elles implique des défis particuliers.[26] Cette collecte s'est ensuite systématisée à d'autres campagnes électorales, nationales ou locales, sous le nom de « Internet en campagne ». Cette première collecte a mobilisé de très importantes ressources humaines (vingt-quatre agents pendant huit mois) disséminées notamment en régions au travers de huit bibliothèques de dépôt légal imprimeur. Ce sont ainsi 63 millions de fichiers (3,4 téraoctets de données) qui ont été collectés,

[26] Pour mesurer la variété de ces collectes et leur richesse, consulter la page : http://www.bnf.fr/fr/collections_et_services/anx_pres/a.collectes_ciblees_arch_internet.html

recouvrant notamment les sites officiels de campagnes, des blogs de particuliers et de « journalistes citoyens » et des sites de militants qui représentent un total de 5'800 sites. (Illien 2008 ; Bonnel, Oury, 2014, p. 8). L'une des collectes projet faisant suite à cette première expérience, celle de la campagne présidentielle de 2007, cristallise encore davantage que la précédente la nécessité de se saisir de tous les nouveaux espaces d'expression collaboratifs présents sur le web, et particulièrement investis par les citoyens lors d'événements politiques majeurs. (Bnf 2006) De toute évidence, la variété des médias sur le web, notamment sociaux, n'est pas sans poser une kyrielle de questions techniques, juridiques et patrimoniales qui doivent être résolues au cas par cas suivant la thématique de la collecte projet.

Nous verrons plus loin à quel point ce type de collecte projet répond à des attentes précises de chercheurs et renouvelle les réflexes documentaires et le sourcing de certains d'entre eux, en proposant des collections avec une unité autant thématique que chronologique. (Bnf 2006) En effet, ce sont uniquement ces collectes projet qui peuvent véritablement donner naissance à une collection dans son acception documentaire.[27]

Enfin, les nouvelles logiques de travail mises en œuvre à l'occasion de ces collectes insistent sur le travail de sélection des différents agents (ou correspondants) chargés désormais de définir : la fréquence et/ou la date à laquelle le robot doit collecter le site, le degré de profondeur et les zones du site à archiver, les potentielles difficultés techniques auxquelles peuvent être confrontés les robots et si une assistance humain est nécessaire pour le suivi du site, notamment en raison de contenus parfois hautement dynamiques. (Illien 2008)

4.2.4 Le Département du Dépôt légal numérique[28]

Comme nous l'avons vu plus haut, la BnF est en charge de collecter les documents au titre de dépôt légal : un département spécifique y est consacré, découpé en cinq départements gestionnaires : le département des documents imprimés et des documents en ligne, le département de l'audiovisuel, le département des estampes et de la photographie, le département de la musique et enfin le département des cartes et plans. La coordination du dépôt légal est prise en charge par le département de l'information bibliographique et numérique et par le Comité de coordination du dépôt légal (CCDL). Des recommandations et des groupes d'expertise sont issus de ce

[27] Entretien téléphonique avec Mme Le Follic, Département du Dépôt légal numérique, Lausanne-Paris, 15 mai 2015.

[28] L'entier de ce chapitre est une synthèse de l'entretien téléphonique avec Mme Le Follic, responsable des collectes au sein du Département du Dépôt légal numérique, le 15 mai 2015.

comité et aident à la bonne marche de la mission du dépôt légal. (BnF 2014c) C'est dans ce contexte qu'Annick Le Follic exerce ses fonctions de responsable des collectes pour le dépôt légal numérique. Au sein d'une équipe relativement restreinte de six professionnels, tous bibliothécaires-système, elle est en charge, depuis le début de la chaîne de production, du suivi des étapes de la collecte. Elle vérifie ainsi les données récoltées et suit de près le travail des robots. Compte tenu des traditions professionnelles plutôt bibliothéconomiques des membres de l'équipe, cette dernière travaille très étroitement avec des informaticiens en charge des dimensions logistiques et techniques de la récolte.

Face à la prolifération des documents hautement interactifs sur le web, le défi réside, pour Annick Le Follic, dans la nécessité de faire évoluer les logiciels-robot pour qu'ils parviennent à capter les mutations documentaires numériques. Bien entendu, la difficulté majeure est celle du tempo rapide des mutations en question. A titre d'exemple, le robot ne parvient pas aujourd'hui à « scroller » les sites monopage comme Twitter ou Facebook qui affichent dynamiquement leurs contenus. Il lui est également impossible de collecter des sites qui exigent, à leur entrée, une date de naissance, le déchiffrement d'une captcha ou le remplissage d'un formulaire. En ce qui concerne les vidéos présentes sur les pages, leur collecte demeure plutôt aléatoire et les sites apparaissent parfois parsemés de cadres noirs, fragmentant les futures archives.

A l'issue des différentes collectes, Annick Le Follic insiste sur le travail d'analyse qui est pratiqué, notamment par l'Observatoire du dépôt légal : synthèses et statistiques sur la production éditoriale nationale composent le rapport annuel[29]. A titre d'exemple, son équipe a pu mettre à jour, suite à la dernière collecte large, des grandes tendances ou types de productions éditoriales de l'Internet français : ainsi, les sites institutionnels (publics ou privés), les sites sociaux, les sites « du quotidien » (petites annonces, météo, etc.) et les sites impliquant le transfert de pratiques culturelles de l'analogique vers le numérique (journaux et livres numériques, jeux, etc) composent les grandes diagonales de l'Internet français en 2014.

Comme évoqué dans l'introduction de cette troisième partie, cet entretien a permis de mettre en exergue la granularité privilégiée dans l'archivage pratiqué par la BnF. En effet, les différents partenaires considèrent l'écosystème du site dans son ensemble et non un/des type(s) de document en particulier. Les préoccupations liées à la volatilité

[29] Le rapport 2013 de l'Observatoire sur le dépôt légal est disponible à cette adresse : http://www.bnf.fr/documents/dl_observatoire_2013.pdf

et à la fragilité des documents issus du web sont permanentes : les bibliothécaires surveillent en continu cette instabilité numérique en accélérant, par exemple, la collecte d'un site qui se renouvelle fréquemment. En identifiant des événements culturels ou politiques de grande envergure ou symptomatiques d'une actualité brûlante, les collaborateurs du projet parviennent à s'en emparer rapidement. Dans cette perspective, de rares collectes dites « d'urgence » peuvent être mises en place, comme cela a été le cas lors des attentats de Charlie Hebdo en janvier 2015 qui ont impliqué, pour le département, une collecte ciblée et renouvelée (jusqu'à quatre fois par jour) des sites de presse, blogs et fils Twitter relayant des informations autour de cet événement tragique. Nous observons bien ici une appréhension holistique des ressources du web dont l'unité archivistique est celle du site Internet. Il est à noter également qu'aucune volumétrie n'a été spécifiquement fixée par département. (Bonnel, Oury, 2014, p. 10)

Les grands défis et chantiers des archives de l'Internet français évoqués par Annick Le Follic sont multiples : deux d'entre eux ont été évoqués lors de l'entretien. L'un d'eux est la mise en place d'une chaîne de dépôt pour se saisir des contenus protégés touchant au livre numérique, spécifiquement les publications au format e-pub publiées par des partenaires identifiés. Ces ressources sont de plus en plus nombreuses et disponibles sur une grande variété de supports. Comme pour les livres imprimés, ces livres numériques sont également soumis au dépôt légal : dans ce contexte, comment les collecter efficacement ? Ce sont les éditeurs qui sont invités à déposer leurs publications numériques, en inversant cette fois-ci la logique du robot moissonneur puisque c'est une chaîne « humaine » qui est mise en place. Le second chantier, qui au moment de notre entretien n'avait pas encore réellement démarré, est celui de l'archivage des applications mobiles. Dans un contexte où l'usage des smartphones se généralise, comment la BnF peut-elle récupérer les applications et de quelle manière ? C'est aujourd'hui un projet complètement inédit qui doit être pensé intégralement.

4.3 Analyse comparative des deux programmes

Dans la dernière partie de ce chapitre, nous allons procéder à une analyse comparative des deux programmes d'archivage du web étudiés aux points 4.1 et 4.2. Nous allons tenter de dégager les différences d'approche ainsi que les points de contact entre les deux institutions. Nous observerons comment le cadre légal influe l'appréhension conceptuelle de l'archivage, les processus de travail et l'accessibilité des archives. Nous mettrons également en lumière certaines similitudes entre ces deux programmes a priori opposés.

4.3.1 Un cadre légal influent et une accessibilité relative

Au cœur du régime différencié des deux programmes qui nous préoccupent réside le cadre légal sur lequel repose largement l'approche mise en œuvre. La nécessité d'une législation pour encadrer les missions d'une bibliothèque patrimoniale apparaît dans les deux cas essentielle. Au-delà de différences techniques entre les législations mobilisées autour des deux projets d'archivage du web, ce sont les approches dont elles sont directement issues qui nous intéressent. Ainsi, l'absence de dépôt légal suisse au niveau national implique pour la BN une approche sélective et thématique du web. Une collecte dite « large », comme pratiquée annuellement par la BnF, serait impossible, aussi bien techniquement que légalement. Comme nous l'avons souligné au travers de l'entretien avec Brigitte Steudler, le travail de réflexion documentaire qui préside à celui de la sélection des sites web est au centre des processus de travail du programme Archives Web Suisse. Même si ce travail de sélection s'inscrivant au sein d'une politique documentaire s'incarne également dans les collectes dites « ciblée » de la BnF, le dépôt légal du numérique français renverse l'approche suisse. En s'emparant indifféremment de la quasi-totalité de la production éditoriale numérique française, la BnF tend à une forme d'exhaustivité sans jugement de valeur documentaire. C'est bien l'armature du dépôt légal qui suppose cette indifférenciation généralisée face aux productions éditoriales numériques. Annick Le Follic souligne cette réalité, qui constitue selon elle une force cardinale du dépôt légal numérique : ne pas préjuger des intérêts futurs des chercheurs. Le dépôt légal interroge également la notion de patrimoine : il ne sanctionne pas symboliquement tel ou tel document puisqu'il les intègre tous sans jugement. Nous sommes donc bien face à deux arsenaux législatifs qui structurent profondément les possibilités de collecte des documents et les processus de travail qui les accompagnent.

Néanmoins, les modes de collectes dites « ciblées » pratiquées par la BnF se rapprochent de celles de la BN : les processus de travail sont, au fond, plus ou moins similaires. En effet, les deux institutions partagent toutes les deux une approche dite « thématique » de l'archivage du web (voir chapitre n°3 « Grandes approches de l'archivage du web et stratégies de collecte »). Des bibliothécaires sélectionnent en amont les sites signifiants et tentent de former des collections parfois thématiques ou gravitant autour d'événements majeurs. Nous observons des outils communs entre les deux structures, comme par exemple le formulaire d'annonce en ligne et l'application BnF Collecte du web (Bcweb) qui occupent la même fonction. Il est également à souligner que les préoccupations liées à la profondeur de l'archivage du site sont également similaires aux deux institutions dans le contexte de ces collectes. La

profondeur particulière de l'archivage pratiquée par la BN différencie les deux institutions. Comme explicité auparavant déjà, la BnF ne pourrait évidemment pas collecter l'entier de son web national en profondeur ; l'exhaustivité théorique du dépôt légal du numérique doit se comprendre ici comme un exercice d'échantillonnage. La BN pour sa part, en sélectionnant précisément la production signifiante et patrimoniale du web suisse, offre une profondeur dans ses collections.

L'accessibilité des archives est également une conséquence directe du cadre législatif différent de chacune des deux bibliothèques : la BnF est obligée d'encadrer son accès pour protéger le droit d'auteur des contenus qu'elle moissonne, alors que la BN est plus souple puisque les accords des producteurs ont été obtenus préalablement. Même si ces accords peuvent permettre une plus grande ouverture des archives, Barbara Signori souligne également les avantages, notamment financiers, du dépôt légal français : « L'archivage se trouve simplifié en termes d'obtentions des droits : lorsque la demande de collecte tombe, on économise des ressources. » (Signori 2015a). Par ailleurs, en raison de la profondeur de l'archivage et de la relative petitesse des collections, les points d'accès, métadonnées, notices de catalogue permettent une meilleure accessibilité des archives helvétiques, selon Annick Le Follic. Cette réalité est également soulignée par Thomas Chaimbault : « [le] choix [d'une approche sélective] permet de collecter une archive de qualité et induit une indexation fine des contenus […] ». (2008). Le travail de description des archives plus abouti de la BN contraste avec la difficulté d'accès aux masses gigantesques offertes par la BnF.

4.3.2 Les retrouvailles internationales

Même si les différences tant conceptuelles que pratiques sont nombreuses entre les deux institutions, il n'en demeure pas moins que l'esprit de collaboration et d'échange reste de mise. En effet, les deux bibliothèques patrimoniales sont membres du Consortium IIPC au travers duquel elles collaborent. C'est notamment au sein du comité de pilotage et des différents groupes de travail que les collaborations peuvent se déployer. A titre d'exemple, les deux institutions s'emparent des mêmes logiciels (notamment le logiciel open source Webspider Heritrix) développés par le Consortium : les expériences de chacune aident à faire évoluer le robot vers plus de performance. Cet engagement international vient rappeler l'impérieuse nécessité de la collaboration dans un contexte d'accélération des mutations du web et des technologies associées.

4.4 Conclusion

Quelles que soient les approches et stratégies choisies par les archives et/ou collections constituées au travers de la variété des programmes actuellement en cours

partout sur le globe, c'est bien la question de leur(s) public(s) qui se pose en filigrane de ces cadres méthodologiques. Si une mémoire patrimoniale du numérique se formalise peu à peu, à qui est-elle réellement destinée ? L'urgence des collectes a-t-elle vraiment laissé le temps de penser le(s) public(s) auxquels elles se destinent ? En effet, si des archives sont conservées, elles le sont toujours en fonction de publics-cibles plus ou moins divers. Les questions cruciales de l'accessibilité à ces nouvelles collections renvoient directement à celles de leurs publics. Nous nous concentrerons au sein du chapitre suivant sur celui des chercheurs et de l'université.

5. Les chercheurs : un public potentiel ?

« Je suis incapable d'avoir la méthodologie qui permettrait d'utiliser de manière scientifiquement valide ce type d'information. Tu as des procédures pour des enquêtes de terrain, t'as des procédures pour des recherches en bibliothèque, mais on n'a pas […] à ma connaissance de procédures pour savoir quoi faire de ces matériaux.

Un chercheur interrogé sur ses appréhensions des collections issues de l'archivage du web de la BnF (Chevallier 2011)

Le public des nouvelles collections issues des différents programmes de l'archivage du web demeure une question centrale. En collectant une masse documentaire particulièrement importante, les différents acteurs de ces programmes cherchent évidemment à la valoriser en pensant notamment son organisation et son accessibilité. Différentes études souvent prospectives sur les représentations, les attentes et les résistances des utilisateurs potentiels de ces archives ont récemment vu le jour. Toutes se préoccupent des besoins du public de ces archives, tant en termes de contenus que de services (Chevallier, Illien, 2011, p. 3), mais également de leurs représentations des archives du web afin « […] d'identifier les moyens permettant d'accroître leur consultation. » (Chevallier, Illien, 2011, p. 3). Même si ces enquêtes et études ouvrent la perspective de considérer des publics variés, nous nous concentrerons dans ce chapitre sur celui des chercheurs. En effet, comme le soulignent Philippe Chevallier et Gildas Illien dans leur étude :

> *« L'usage du web est aujourd'hui omniprésent pour les chercheurs en sciences humaines, […] comme possibilité d'accéder à de la documentation scientifique, mais également comme terrain de recherche […] le web devient un lieu d'exposition du chercheur lui-même, désormais actif dans [sic] les réseaux sociaux et sur les blogs. »* *(Chevallier, Illien, 2011, p. 7)*

S'adresser au public des chercheurs est également nécessaire au sens où ces derniers ont souvent peu l'expertise technique pour s'emparer seuls de ces ressources :

> *« A unique and emerging use of archives is as a research service for scholars. Very few academics, especially in the social sciences and humanities, have the computational expertise or resources to crawl and download large portions of the web for research. »* *(Leetaru 2012)*

Dans ce chapitre, nous identifierons les champs du savoir les plus concernés par la mobilisation des archives du web : certaines tendances se dessinent déjà. Nous pointerons également la nécessité pour les programmes d'archivage du web de multiplier les collaborations avec la communauté académique afin d'offrir des contenus appropriés, une accessibilité et des interfaces qui placent l'utilisateur au centre.

5.1 Introduction : le cas de « l'Internet en campagne »[30]

Les collectes projet organisées par la BnF dès 2002 pour couvrir différentes élections françaises (présidentielles et législatives) demeurent sans doute, à ce jour, le cas d'école le plus probant pour exemplifier la force et l'impact des initiatives de l'archivage du web sur la communauté des chercheurs. L'histoire politique et sociale de la France ainsi que sa sociologie électorale se trouvent ainsi enrichies des « […] matériaux web diffusés par les candidats, les partis politiques, les médias […] » (Greffet, 2012, p. 80). La chercheuse Fabienne Greffet contextualise ce phénomène récent du « web politique » et souligne la pertinence de telles collectes :

> *« Les sciences sociales, et parmi elles la science politique, apparaissent également travaillées par ces transformations, accompagnant […] la généralisation de l'Internet. En s'en tenant au web en démocratie […], on constate que celui-ci ouvre de nouveaux terrains et de nouveaux enjeux pour la recherche en science politique. » (2012, p. 79)*

Dans le contexte de l'écosystème des sites liés aux campagnes électorales, nous retrouvons bien la dimension éphémère des contenus du web, rattachés à un événement circonscrit, qui sont voués à disparaître une fois les élections achevées : la nécessité de leur collecte apparaît ici dans toute son urgence. Même si ces collectes peuvent sembler parfois « […] trop circonscrites à des producteurs institutionnels […] » (Greffet, 2012, p. 80), elles n'en demeurent pas moins plébiscitées par les archinautes de ces dernières années. Les sites d'organisations politiques et des candidats, les blogs de militants, les observatoires et sites d'information sur les élections sont autant de lieux où s'élaborent les logiques et les rouages d'un processus électoral. Ce sont également des documents qui se déplacent désormais de l'analogique vers le numérique. Compilés et traités, ils permettent, tant au niveau local que national et dans un souci de neutralité des sensibilités politiques, de décrire les candidats et leurs organisations et de saisir les regards et opinions sur la campagne (BnF 2006). Les caricatures pendant les élections présidentielles de 2002 et 2007, la critique du traitement télévisuel de l'insécurité pendant la campagne électorale de 2002 ou la parole des femmes candidates aux élections sont autant de sujets traités par des étudiants ou des doctorants en sciences politiques, sociologie et linguistique qui ont consulté les archives du web pour leurs travaux (Aubry et al. 2008, p. 14).

Ces projets de collecte constituent également un modèle puisqu'ils sont l'occasion de formaliser des partenariats fructueux avec le monde universitaire. Ce type de

[30] L'expression « Internet en campagne » a été utilisée par la BnF dans son dossier de presse sur l'archivage des sites électoraux. (BnF 2006)

collaboration est en permanence pointé comme une nécessité pour la constitution de collections pertinentes :

> *« […] the most important things to do is to engage domain researchers with expertise not just in Internet research, but in fields such as sociology, political science, other social sciences, physics and other sciences, the arts & humanities, and others as these infrastrucutres are designed so that the needs of national researchers are reflected in the collections created. » (Meyer, Thomas, Schroeder, 2011, p. 7)*

La BnF s'est ici associée avec la Fondation Nationale des Sciences Politiques (bibliothèque-pôle associé à Sciences Po) afin, notamment, de mener des tests d'utilisation des interfaces d'accès aux collections. Sciences Po était également responsable de sélectionner des chercheurs ainsi que de suivre et valoriser leurs activités en lien avec la constitution de ces archives. Mesurer les attentes des chercheurs, leurs besoins et leurs réticences constitue un défi majeur qui se généralise quel que soit le domaine, afin de confirmer que :

> *« […] les analyses générées par l'observation de ce qui intervient en ligne contribuent à interroger, reformuler voire redéfinir certains modèles ou concepts de science politique, […] forgés […] en dehors des questionnements sur les technologies de l'information et de la communication. » (Greffet, 2012, p. 83)*

5.2 Attentes et représentations des chercheurs

Annick Le Follic précise, lors de notre entretien mentionné plus haut, que ce sont surtout les chercheurs issus des sciences sociales, politiques et de l'art numérique qui composent aujourd'hui le public privilégié des archives du web de la BnF. Claude Mussou, responsable du dépôt légal du web des collections de l'INA, y ajoute encore d'autres disciplines : media/cultural studies, anthropologie et ingénierie. (2012) Ce public pluridisciplinaire s'intéresse notamment à comparer l'état du « web vivant » avec les archives constituées. C'est également, au-delà des contenus informationnels présents sur les sites, la notoriété et le référencement de ces derniers qui préoccupent certains chercheurs. Les modalités d'archivage permettent aujourd'hui de mettre en lumière, notamment au moyen de « cartographies dynamiques », les relations qu'entretient un site précis avec son contexte : « […] circuits d'influence, réseaux d'appartenance, d'affiliation, d'opposition ou, plus généralement, relation entre les acteurs. » (BnF 2006). Les leviers d'analyse pour faire parler ces archives sont nombreux : « […] grilles de description des contenus, analyses de discours, analyses de l'hypertextualité et cartographie des espaces web et des liens, analyse sémiotique… […]. » (Greffet, 2012, p. 81).

Le cas des historiens face aux nouvelles sources que constituent les archives du web cristallise à la fois des attentes spécifiques et des changements de paradigme. Le web social forme, par exemple, « de nouvelles archives sur la vie quotidienne » (Joutard, 2013, p. 146) incontournables pour comprendre les phénomènes de notre époque. La collecte « (S') écrire en ligne : journaux personnels et littéraires »[31] menée par la BnF illustre parfaitement cette volonté de mettre en lumière de nouvelles pratiques du *blogging* nées du passage du papier à l'écran. Ces nouvelles traces, extrêmement nombreuses, viennent enrichir le « contrat de vérité » entre l'historien et son lecteur. Même si l'historien Philippe Joutard affirme que « […] les historiens n'ont plus le choix et doivent tous travailler dans le cadre du numérique. » (2013, p. 149), il n'en demeure pas moins que de nombreuses questions méthodologiques et épistémologiques sont toujours en discussion aujourd'hui, comme en témoigne l'abstract d'une contribution récente à un colloque d'historiens : « En quoi est-ce que les nouveaux documents numériques peuvent-ils constituer de nouvelles sources pour les historiens ? A quel nouveau type d'archives le chercheur en histoire doit-il se confronter ? », mais aussi : « Qu'est-ce que la source numérique dit mal, ou ne dit pas du tout et quels recours méthodologiques sont possibles pour l'historien ? » (Abbou 2013). Certains professionnels et historiens sont traversés par des interrogations sémantiques et cherchent tout d'abord à qualifier ces nouveaux contenus : s'agit-il de traces, d'indices, de ressources ou encore de simples données ? Aucune dénomination ne semble réconcilier les uns et les autres : Marc Bloch semble confondre traces et documents : « […] qu'entendons-nous en effet par « documenta » sinon une trace, c'est- à-dire la marque, perceptible aux sens, qu'a laissée un phénomène lui-même impossible à saisir »[32] (Bloch 1999), alors que Claude Mussou conclut :

> *« Cette inflation et massification des contenus publiés, leurs fragmentations, propagation et dématérialisation engagent d'ailleurs une rupture épistémologique à la fois pour l'archiviste et le chercheur qui considèrent désormais des « ressources » ou des « données » davantage que des documents. » (2012)*

« […] la dissociation entre le support et son contenu […] a entraîné la disparition de la notion d'original et, in fine, celle de son support qui fondait les principes et pratiques de l'archivistique. » (Mussou 2012). Le particularisme des documents du web entraîne une redéfinition des concepts traditionnels de l'archive et invite à une nouvelle utilisation de ces ressources. Si la volatilité intrinsèque de ces contenus peut pousser certains historiens à s'en emparer, il demeure délicat pour eux d'appréhender des « corpus » de données qui n'en sont pas vraiment : l'enjeu réside bien, pour les

[31] Pour davantage d'informations sur cette collection, consulter : http://blog.bnf.fr/lecteurs/index.php/2009/04/du-journal-intime-au-blog/

[32] Cité par Claude Mussou dans son article (Mussou 2012).

administrateurs des programmes d'archivage du web, dans l'organisation des contenus en collaboration avec les pôles de compétence concernés : « […] travailler en relation étroite avec la communauté académique [est une] priorité pour la constitution de collections qui […] s'efforcent de répondre le mieux possible aux exigences de recherche. » (Mussou 2012). Le processus de légitimation de ces nouveaux corpus pourrait provenir aussi de l'institution bibliothéconomique, comme le suggère Gildas Illien : « Parce que l'archive du web devient plus intelligible et plus respectable en tant qu'objet bibliothéconomique, on commence à la regarder comme une *collection* à part entière. » (2011, p. 67).

Annick Le Follic nous explique lors de notre entretien qu'il existe en France une vraie volonté politique de créer des espaces de recherche sur des sujets majeurs au travers de la mise en place de laboratoires d'excellence, contractés « Labex » ; ce sont les promoteurs de grandes thématiques de recherche, choisies par certaines instances gouvernementales. Certains d'entre eux, à l'instar du labex EHENE[33], travaillent par exemple sur la Première Guerre mondiale et réfléchissent sur la façon dont le numérique peut aider à penser autrement les travaux des historiens sur cette tranche de l'histoire contemporaine. La BnF participe aux projets de ce laboratoire dont l'une des pistes de recherche est la mise en place d'outils à destination des chercheurs, qui leur permettraient de mieux investir les collections et les archives proposées par la bibliothèque. En collaborant directement avec les chercheurs concernés, la bibliothèque participe à la création de leviers spécifiques pour mieux appréhender cette masse documentaire nouvelle que constituent les archives du web. De plus, la BnF multiplie les collaborations avec la communauté scientifique lors de conférences et autres interventions qui ont pour but d'expliquer aux chercheurs ce que sont les archives du web. C'est également l'occasion de mener des enquêtes sur les attentes des universitaires en termes d'interfaces d'accès. Par ailleurs, la multiplication des collaborations avec le monde scientifique peut avoir des répercussions budgétaires positives pour les programmes d'archivage du web, trop souvent méconnus ou incompris :

> *« [...] partnering with researchers may help: there are only a few sources of funding for the creation and operation of web archives compared with the myriad funding opportunities for research. [...] it directly demonstrates the value of those archives to new audiences and disciplines that may be able to partner with those archives on proposals, potentially offering new funding opportunities.»*
>
> *(Leetaru 2015)*

[33] Pour le détail des programmes de ce laboratoire, consulter : http://www.labex-ehne.fr/

Les auteurs de l'étude menée par la BnF en 2011 auprès, notamment, des chercheurs, soulignent plusieurs résultats. Les premiers concernent l'appréhension globale du web et de ses usages : Internet est perçu comme un terrain fécond pour « prendre le pouls » d'un sujet : « J'étais allé voir un peu ce qu'on racontait », déclare l'un des chercheurs interrogés. (Chevallier, Illien, 2011, p. 7) C'est également le lieu de l'innovation et des « nouveaux objets » à dénicher. Certains chercheurs pratiquent une veille spécifique sur un sujet. Dans cette logique, d'autres ont organisé des répertoires de sites qui constituent ainsi de petites archives personnelles autour de thématiques choisies et y stockent d'importants volumes de données. La consultation d'Internet Archive semble très courante même si certains chercheurs pointent les trop nombreux liens morts. (Chevallier, Illien, 2011, p. 7-9)

De façon globale, la nécessité et l'urgence d'organiser une « mémoire du web » sont partagées par tous dans la communauté scientifique. La volatilité, la fragilité et la plasticité du web sont autant d'éléments constitutifs de la toile qui génèrent des attentes d'archivage très claires de la part des chercheurs : « Il y a quand même une espèce de crainte de la disparition ; on voit des choses disparaître. » (Chevallier, Illien, 2011, p. 11) confie l'un d'eux. Les chercheurs semblent ainsi prendre conscience que l' « On recense les premiers cas de disparition critique des contenus de l'Internet et les premières demandes de particuliers, d'entreprises et même de partis politiques qui ont perdu leurs données. » (Illien, 2011, p. 67). Archiver le web semble être le moyen privilégié pour souligner la dimension mouvante d'Internet : si une page est amenée à se transformer sans cesse, son archivage peut être le levier préféré qui inscrira le site dans une perspective historique. Dans ce contexte, l'étude de la BnF observe une tension : tout document peut recéler un certain intérêt, mais le tri, la sélection et les choix sont très attendus par les chercheurs. (Chevallier, Illien, 2011, p. 9)

5.3 Interrogations et résistances des chercheurs

De toute évidence, la mise en place des archives du web n'est pas sans poser un catalogue d'interrogations, voire de résistances diverses, à la fois méthodologiques, éthiques et pratiques. En amont de la mobilisation d'archives du web dans un travail scientifique, les chercheurs soulignent leur incertitude : un site Internet peut-il réellement constituer une source fiable ? Quelle procédure existerait-il pour valider la qualité d'une telle source ? Comment justifier le choix de convoquer tel site plutôt qu'un autre ? L'établissement de collections cohérentes peut générer un capital légitime : « […] pour que [l'usage du site] soit justifié, il faudrait que l'archive citée puisse être inscrite dans un corpus aux contours maîtrisés et partagés par une communauté de chercheurs. » (Chevallier, Illien, 2011, p. 10). Comme nous l'avons vu plus haut,

l'exhaustivité supposée de ces « corpus légitimes » est impossible à atteindre ; cette réalité constitue également un frein à leur usage pour les chercheurs : « […] la base documentaire n'est pas jugée assez solide pour permettre un travail académique contrôlable […] » (Chevallier, Illien, 2011, p. 10). Un protocole de validation peut venir palier ce biais intrinsèque à la mobilisation de ces archives, mais le sentiment de manque de fiabilité reste prégnant. Néanmoins, la mise en place d'une politique documentaire et donc d'une sélection est très attendue par les chercheurs. La fonction et les dispositifs de destruction inhérents à la sélection de l'archive sont primordiaux dans ce contexte presque borgésien. (Chevallier, Illien, 2011, p. 10)

Même si Annick Le Follic note un intérêt et une curiosité de la part des chercheurs, elle confirme également les résistances décrites en insistant davantage sur les appréhensions techniques du public académique. En effet, au-delà des questions qui entourent la pertinence des sources, les chercheurs sont peu habitués à interagir avec les interfaces actuellement en place : ils doivent aujourd'hui inventer leur propre manière de conduire leur recherche. De plus, les conditions d'accès peuvent en rebuter certains : la nécessité de devoir se déplacer au sein même des bibliothèques[34] donnant accès aux collections constitue, selon Annick Le Follic, la contrainte principale. Avec environ une trentaine de consultations mensuelles, les archives du web français se doivent de réfléchir aux points de contact entre leurs collections et leurs publics. Toujours en termes d'accessibilité aux archives, mais cette fois-ci concernant spécifiquement les OPAC et les autres interfaces, l'indexation plein texte (*full text*) des collections demeure la voie d'accès privilégiée. Claude Mussou souligne une dimension politique de cette forme d'indexation qui prend tout son sens dans le milieu scientifique et « neutre » de la recherche :

> *« Quand elle est disponible, l'indexation « full text » du web archivé assure […] une neutralité des réponses et, par là même, offre une garantie au chercheur en même temps qu'un contrepoids à la substitution de la mémoire collective par les monopoles industriels du web mondial. » (2012)*

Brigitte Steudler, quant à elle, note aussi une frilosité certaine de la part du public universitaire. Il s'agirait selon elle d'attendre quelques générations pour consolider des collaborations installées entre l'université et le programme d'archivage de la BN, même si elle reconnaît un certain dynamisme du côté des sciences dures dans les EPF. Tout comme Annick Le Follic, elle pointe la difficulté d'accéder aux collections pour le public qui doit également se déplacer en bibliothèque. Les difficultés de prise en main des

[34] Il est à noter que trois autres bibliothèques en région donnent désormais accès aux Archives de l'Internet : Montpellier, Strasbourg et Nancy. A l'avenir, d'autres institutions devraient également offrir cet accès.

interfaces constituent également un frein quand elles ne concernent pas aussi le personnel de l'institution.[35] Ce témoignage corrobore les observations d'une autre étude qui pointe l'importance de proposer des interfaces calibrées pour les chercheurs :

> *User interfaces as a whole today are designed for casual browsing by non-expert users, with simplicity and ease of use as their core principles. As archives become a growing source for scholarly research, archives must address several key areas of need in supporting these more advanced users. (Leetaru 2012)*

Les interrogations se situent également au niveau du périmètre de la collecte : même si la plupart des chercheurs s'accordent autour du bien-fondé de l'archivage des sites institutionnels et des blogs, les actions ou traces individuelles (actes d'achat, conversations, etc.) laissées sur le web sont considérées avec davantage de circonspection. Les questions liées aux droits individuels, à l'image et aux espaces privés/publics émergent rapidement avec leurs lots d'inquiétudes largement partagées quant à la potentielle exploitation des données personnelles. Lorsque le périmètre est défini, notamment par un outil comme la charte documentaire, les chercheurs s'interrogent sur les processus de « hiérarchisation » et de « discrimination » des documents. A nouveau, il s'agit de repenser les grands ensembles classificatoires traditionnellement établis par les bibliothèques et d'intégrer des notions de lignées, maillons et hiérarchies nouvelles. (Chevallier, Illien, 2011, p. 17)

5.4 Conclusion

Comme nous l'avons décrit dans ce chapitre, les attentes, résistances et autres interrogations des chercheurs sont variées et multiples, se mêlant parfois les unes avec les autres. La pluralité des disciplines auxquelles est rattaché le public des chercheurs, qui ne se cantonne pas à celui des sciences humaines, ne facilite pas un bilan global du regard, encore émergent, que porte le milieu universitaire sur l'archivage du web. Nous pouvons néanmoins observer certaines récurrences : les chercheurs pointent ainsi la nécessité de conserver de nouvelles formes d'expressions numériques en plus des sites traditionnels comme les blogs. Cette conservation doit s'accompagner d'une reconceptualisation des modèles traditionnels de l'archivage. L'organisation, la hiérarchisation, voire la discrimination des contenus issus du web sont attendues pour appréhender au mieux les nouveaux corpus. Les pertes de certains contenus nés numériques et les liens morts inquiètent certains acteurs du monde académique, qui désormais s'accordent à dire qu'une mémoire du web doit être

[35] Entretien avec Mme Brigitte Steudler, responsable de la Documentation vaudoise, personne de contact opérationnel d'Archives Web Suisse pour Vaud au sein de la Bibliothèque cantonale et universitaire de Lausanne (BCU), Lausanne, 16 avril 2015.

constituée. Les doutes et les interrogations se cristallisent autour de la fiabilité de ces nouvelles archives dont les contours documentaires peinent à être scientifiquement définis ; en filigrane se pose la question de la légitimation du statut de collection de ces archives. La bibliothèque comme « tiers de confiance » (Chevallier, Illien, 2011, p. 3) peut avoir un rôle à jouer dans ce processus. Plus concrètement encore, ce sont les difficultés d'accès, autant physiques que techniques, qui préoccupent les chercheurs : le déplacement supposé au sein des bibliothèques et les interfaces difficiles à maîtriser empêchent trop souvent le public de s'approprier ces nouvelles ressources. Enfin, l'instabilité du média Internet, la volatilité des données et la difficulté à traiter de gros volumes de données souvent très hétérogènes (Mussou 2012) constituent les principaux freins méthodologiques rencontrés par le monde de la recherche.

Certains de ces écueils trouvent aujourd'hui des solutions au travers de groupes d'études qui réfléchissent aux futurs de l'archivage du web. Les innovations technologiques peuvent apporter des réponses concrètes aux attentes et problèmes évoqués dans ce chapitre. La collaboration internationale entre les différents acteurs de l'archivage du web est notamment le lieu d'échanges et d'élaboration de nouvelles « normes [et logiciels] pour la collecte, la préservation et l'accès à long terme aux contenus de l'Internet » (Illien, 2011, p. 63).

6. Recommandations

Ce chapitre, résolument tourné vers l'avenir des archives du web, se propose de déployer un panorama non-exhaustif des futurs possibles de l'archivage du web. La littérature prospective sur cette question est peu nombreuse : beaucoup de sources soulèvent davantage des interrogations et des problèmes rencontrés sur le terrain de chacun des différents programmes d'archivage. Comme évoqué précédemment, le lieu des innovations en matière de conservation du web se situe surtout dans le cadre de collaborations internationales. Nous étudierons ainsi certaines pistes d'innovation proposées par trois études, toutes répertoriées par le consortium IIPC. Nous observerons également la façon dont ces pistes renvoient parfois aux préoccupations concernant les cas que nous avons étudiés. Les résonnances apparaîtront également avec le public des chercheurs à qui les innovations décrites sont parfois destinées. La création d'interfaces et de voies d'accès aux collections, la description des fonds d'archive et la documentation de leur contexte ou encore la possible fonction d'authentification des archives constituent quelques-unes des pistes récurrentes.

6.1 Le consortium IIPC, un laboratoire des futurs de l'archivage du web

Les sources concernant l'archivage du web sont plutôt disparates. En effet, le peu de recul que les différents acteurs peuvent avoir sur ce type d'initiative induisent une relative pauvreté des textes sur le sujet. Les publications qui ont été mobilisées tout au long de ce travail concernent spécifiquement les programmes étudiés ou s'y rattachent fortement. Beaucoup de sources, la plupart du temps en anglais, demeurent souvent extrêmement techniques et cherchent à répondre à des problèmes spécifiques rencontrés par tel ou tel programme. Plusieurs bibliothèques proposent des bibliographies générales et sélectives d'ouvrages et d'articles sur la question. L'ancienneté de beaucoup d'entre elles pose parfois problème ; les questions qui y sont développées sont quelquefois caduques aujourd'hui. Néanmoins, d'autres sources générales, même si parfois anciennes, nous ont permis d'identifier des enjeux profonds qui restent toujours d'actualité. Les publications scientifiques sont souvent produites par des chercheurs ou spécialistes en sciences de l'information et de la communication. Les historiens, traditionnellement concernés par les enjeux de la mémoire, sont également des contributeurs réguliers de cette littérature. Au fond, le caractère totalisant de l'archivage du web peut être l'occasion pour chaque discipline de s'engager dans une publication mettant en lumière ses propres enjeux avec les programmes, « […] chaque [discipline] mobilisant des constitutions de corpus et outils d'analyse différents. » (Bonnel, Oury, 2014, p. 10).

Pour éclairer les perspectives futures de l'archivage du web, nous nous sommes principalement saisis des publications offertes par le consortium international pour la préservation de l'Internet (IIPC). Cet organisme regroupe aujourd'hui une quarantaine d'institutions du monde entier : bibliothèques, centres d'archives, universités, etc. Il concentre son action autour de trois axes : « le développement logiciel, la communication et le lobbying et la création collaborative de contenus » (Illien, 2011, p. 61). Ce sont, à l'origine, une dizaine de bibliothèques nationales européennes et nord-américaines qui s'associent à l'organisme Internet Archive pour fonder le consortium en juillet 2003. Rapidement, les premiers membres œuvrent à l'établissement « […] d'une communauté de formats, de normes et de logiciels, facilitant la production de données homogènes par les institutions. » (Illien, 2011, p. 62). Les missions fondamentales de l'IIPC sont le développement de solutions pour sélectionner, collecter, préserver et rendre accessible les contenus du web, « faciliter la couverture internationale des collections d'archives […] » du web et enfin, « plaider au niveau international en faveur d'initiatives […] » pour la préservation de la mémoire du web. (Illien, 2011, p. 63) Afin d'y parvenir, le consortium s'engage à favoriser les échanges entre les membres, le développement de logiciels adaptés (le plus souvent libres), l'organisation de conférences et d'ateliers pour sensibiliser aux questions liées à l'archivage du web. (Illien, 2011, p. 63) Ainsi, les trois axes de travail du consortium rejoignent ceux associés au circuit du document en bibliothèque : collecte, consultation et préservation. A titre d'exemple, les travaux de normalisation par l'ISO d'un format conteneur des archives du web (WARC), chapeautés par l'IIPC, ont largement participé à la légitimation de ces nouvelles archives. La condition transnationale du web implique une kyrielle de questions de « géopolitique patrimoniale » (Illien, 2011, p. 67) auxquelles seul un organisme international comme l'IIPC peut tenter de répondre.

L'IIPC cherche à « améliorer la sensibilisation aux questions liées à la préservation des contenus de l'internet et aux initiatives associées, notamment par le biais de [...] publications. » (Illien, 2011, p. 63). En 2011 a été publié un rapport des chercheurs de l' « Oxford Internet Institute »[36] sur les futurs possibles des archives du web dont nous présenterons plus bas les résultats. Un article sur le rôle et le futur des archives du

[36] Pour davantage d'informations sur cet institut, consulter : http://www.oii.ox.ac.uk/

web et des études de cas, toujours publié dans le cadre de l'IIPC, viendront compléter et illustrer cette première étude et offrir ainsi le panorama de nos recommandations.

6.2 L'étude de l'Oxford Internet Institute

6.2.1 Scénarii d'experts

Dans leur étude sur les futurs du web, les experts de l' « Oxford Internet Institute » identifient plusieurs scénarii possibles pour l'archivage du web au niveau mondial. Deux d'entre eux ont l'intérêt de soulever des questions importantes. Le scénario nommé « Apocalypse » annonce l'avenir le plus sombre aux programmes d'archivage. Ainsi, les archivistes du web risquent d'être incapables de suivre le rythme particulièrement soutenu des évolutions techniques du web, ainsi que les changements constants de formats de fichier qui entraîneront, à terme, une illisibilité totale des contenus archivés. Les archivistes seront ainsi face à un échec permanent de l'indexation des collections et des technologies de recherche d'une masse documentaire toujours plus volumineuse. Une attitude démissionnaire frappera les archivistes qui opteront finalement pour le « Let Google do it ». L'impossibilité de prendre en charge les archives du web ne manquera pas de laisser penser aux générations futures que le web fut un épisode anecdotique de l'histoire de l'humanité. (Meyer, Thomas, Schroeder, 2011, p. 6)

Le second scénario intitulé « Dusty Archive », moins alarmiste que le premier, expose également quelques dangers probables. L'un d'eux consisterait à ce que le public considère le web vivant comme des archives permanentes qui se constituent en continu, sans chercher à recourir à des contenus pérennes et organisés. Le désintérêt des chercheurs observé par les auteurs de l'étude, ainsi que l'absence de technologie adéquate pour accéder aux archives déjà existantes expliquent ce potentiel danger. Les archivistes ont ici leur rôle à jouer pour éviter cet écueil en valorisant les programmes d'archive et en développant des outils appropriés. Cette batterie d'outils d'investigation des collections peut constituer un levier majeur pour inscrire les archives du web comme nouveaux réflexes documentaires et intellectuels. (Meyer, Thomas, Schroeder, 2011, p. 7)

6.2.2 « Apprendre du web vivant »

Des outils de pointe existent aujourd'hui pour appréhender, étudier et investir le web vivant. Plutôt que de nécessairement conceptualiser des outils dédiés, les membres de l' « Oxford Internet Institute » proposent de transposer certains de ces outils aux archives du web. Ces pistes sont autant de mises en valeur possibles des collections

qui pourraient être analysées et exploitées afin d'offrir une réelle plus-value aux services d'archive du web. Nous en rapportons ici quatre d'entre elles :

La ***visualisation*** peut constituer une fenêtre d'accès inédite aux archives. Dans l'esprit des infographies, elle permettrait de visualiser la façon dont les différentes archives sont reliées entre elles. Un fort développement de cet outil pour le web vivant existe déjà. La ***recherche profonde*** permet d'interroger finement de gros ensembles de données. La prolifération des informations postées (puis archivées) exigerait ainsi de nouveaux moyens d'accès à de très gros volumes d'informations. L'***analyse des réseaux sociaux*** (« Social Networks Analysis » (SNA)) n'a pas été adaptée aux archives. Ces outils d'analyse spécifique pourraient permettre aux archivistes du web l'analyse des liens hypertextes comme révélateur de la structure des interactions des différents sites web composant leurs collections. Les liens et leur analyse disent quelque chose de la nature du réseau. Enfin, cette analyse pourrait être complétée par ***l'archivage de tous les liens et autres annotations*** (favoris, signets) qui pointent vers les sites archivés et observer leurs évolutions dans le temps. (Meyer, Thomas, Schroeder, 2011, p. 9-12)

D'autres outils, parfois très techniques, sont proposés par l'étude : capture des interactions sociales et des comportements numériques, extraction des données géographiques des archives pour réaliser des cartographies montrant l'évolution dans le temps d'un phénomène, étudier les usages du web (et non plus seulement les contenus), généraliser la pratique du web sémantique, etc. (Meyer, Thomas, Schroeder, 2011, p.13-16)

6.2.3 Des futurs et des défis

Les défis qui accompagnent les futurs de l'archivage du web sont nombreux et variés. Des pistes d'innovations, cette fois-ci propres aux archives du web et à leur traitement possible, se déclinent au sein du rapport de l' « Oxford Internet Institut ». Nous en présentons ici seulement quelques-unes[37].

La première piste est celle dite du ***web cumulatif*** : il s'agit de considérer le web archivé littéralement *en parallèle* du web vivant. Cette organisation en filigrane, de couches d'archives, viendrait combler la fragmentation et les trous du web (comme les liens morts qui désormais pointeraient vers la ressource archivée). Cette piste

[37] Notons que les leviers techniques et informatiques qui pourraient permettre de mener à bien ces différents défis sont encore balbutiants ou à l'état de prototype. Nous renvoyons le lecteur à l'étude pour le détail technique des innovations décrites dans ce chapitre. Le lecteur trouvera également des réponses techniques à ces questions au sein du rapport « Web Archiving Use Cases » (Reynolds, 2013).

supposerait un changement structurel et profond du web ; il s'agit d'une piste pour le moins révolutionnaire qui demeure aujourd'hui relativement utopique.

S'il est aujourd'hui possible de comprendre l'organisation et les usages des sites présents sur le web et de consulter certains d'entre eux qui n'existent plus, il demeure impossible de comprendre ***l'usage passé des archives*** du web. Afin d'y parvenir, il s'agirait d'archiver également les journaux des serveurs (« servers logs ») des sites d'archivage du web ; de cette façon, il deviendrait possible de comprendre et d'étudier comment les archives du web ont été ou sont utilisées. Le défi consisterait donc à mettre en place une infrastructure qui permettrait non seulement de voir comment le web était avant sa disparition, mais également les usages associés à ce web dans le passé. Certains chercheurs expriment spécifiquement cette attente : « L'Internet, c'est d'abord une pratique. Donc : archivons la pratique [notamment des archives], parce que sinon on va perdre la pratique. » (Chevallier, Illien, 2011, p. 16).

Un usage possible des archives du web est celui de ses ***images*** et de son fort potentiel visuel. En effet, il est possible de saisir certains changements du monde au travers des images circulant sur la toile. En extrayant sur une certaine durée des images d'archives (par exemple depuis la plateforme « Flickr ») d'un même bâtiment, cela permettrait, au-delà de la simple comparaison entre l'ensemble des clichés recueillis, de superposer les images et de proposer ainsi un rendu visuel de l'évolution du bâtiment. L'exercice pourrait se décliner avec n'importe quel sujet photographié.

L'***exploitation statistique*** des archives du web constitue également une opportunité majeure. Quels sont les outils d'analyse à mettre en place pour faire parler de grandes collections d'archives du web ? Comment ces outils statistiques permettraient de mieux comprendre la structure des collections et conséquemment celle du web en général ? En s'intéressant, par exemple, aux langues des sites web ou à leur date de création, il serait possible de dégager des grandes tendances structurelles du web. C'est dans cette perspective que s'inscrivent les travaux d'analyse menés par l'Observatoire du dépôt légal de la BnF : la collecte large est ainsi analysée en proposant, parmi d'autres, des statistiques liées à la variété des types de fichiers récoltés.

Toujours en lien avec l'analyse structurelle du web, les auteurs de l'étude proposent de réfléchir à la façon de rendre compte de la ***prolifération d'une idée sur le web***, sa viralité et ses déplacements. Pour repérer et comprendre où les idées surgissent et comment elles se propagent sur le web, il s'agit de pouvoir remonter à l'origine de l'idée. Cette archéologie suppose une profondeur et une granularité des archives très importantes. La temporalité du web, c'est à dire le tempo des publications et les

hyperliens qui les relient, doit être archivée et analysée. Sans une profondeur suffisante de l'archivage, cette dimension est impossible à extraire des archives. La chercheuse Emily Reynolds propose certains exemples de projets qui mobilisent ces outils, notamment « Babel 2012 Web Language Connections »[38]. Ce projet hollandais éclaire l'usage des langues des internautes (notamment le bilinguisme) et propose une visualisation des « liens inter-langues » (« inter-language links ») entre les sites étudiés. (2013, p. 4)

Enfin, la question du ***web illicite*** est également abordée par les auteurs de l'étude qui s'interrogent sur la meilleure façon de rendre compte des matériaux illicites circulant sur le web. Les contenus sexuels illicites, sur les drogues, sur les groupes prônant la haine raciale, le terrorisme, etc. sont nombreux. Quelle entité serait habilitée à prendre en charge leur archivage et dans quel cadre juridique ? Ce genre d'archive pourrait autant intéresser les chercheurs que certaines autorités, la justice ou encore les professionnels de la santé. L'enjeu réside bien ici dans la mise en place d'un mécanisme juridique pour protéger et légitimer l'institution garante de ces documents, qui saurait mettre en valeur leur intérêt scientifique.

A l'issue de leur étude, les auteurs pointent quelques constats : l'absence actuelle d'interfaces stables et conviviales pour construire des archives du web et les analyser empêche certains programmes de se déployer. Les mêmes interfaces découragent plus d'un usager et les institutions manquent trop souvent de ressources financières. Les auteurs encouragent les collaborations entre les différents acteurs des programmes : techniciens, informaticiens, chercheurs et bibliothécaires doivent travailler de concert pour résoudre les défis décrits. L'organisation d'un « hackathon » permettrait de mobiliser les programmeurs autour de solutions novatrices et créatrices pour de nouveaux outils et interfaces. (Meyer, Thomas, Schroeder, 2011, p.17-25)

6.3 L'étude de Kalev Leetaru

6.3.1 Interfaces et voies d'accès aux archives

Comme nous l'avons déjà évoqué au cours de ce travail, l'étude de Leetaru insiste sur l'opportunité de mettre en place des **interfaces d'accès** aux archives les plus efficaces possibles. Les archives du web étant amenées à s'enrichir en permanence, l'auteur insiste sur la création d'interfaces qui sauraient explorer de très gros volumes de données : « New programming interfaces and access policies are needed to enable this new generation of schloarship using web archives. » (Leetaru 2012). L'interface

[38] Pour davantage d'informations sur ce projet, consulter : https://github.com/norvigaward/2012-naward25/wiki/Babel-2012---Web-Language-Connections

bien connue du réseau social Twitter pourrait, selon l'auteur, constituer un modèle standardisé très simple d'utilisation : « If archives took the same approach with a standardized interface like Twitter's, researchers could leverage these huge ecosystems for the study of the web itself. » (Leetaru 2012). Il s'agit également, pour les concepteurs des futures interfaces d'accès, de penser spécifiquement au public-cible des chercheurs, qui formera sans doute une communauté importante se saisissant des futures archives du web. Au sein de ces interfaces, une batterie d'outils devrait être présente pour investir au mieux les collections et offrir ainsi un maximum de visibilité aux contenus agrégés.

Comme soulevé précédemment par le cas de la BnF, l'indexation des collections, ou du moins, l'inventaire de celles-ci sont des voies d'accès possibles pour mieux rechercher au sein des archives. Si l'indexation reste souvent difficile ou trop coûteuse, la **description fine des archives** au travers de métadonnées variées constituerait également une mise en valeur des fonds. Cette pratique suppose que les administrateurs des programmes connaissent précisément le contenu de leurs archives, ce qui n'est pas toujours le cas. Des métamoteurs de recherche pourraient également voir le jour : sur le modèle du métamoteur bibliographique « WorldCat », des recherches fédérées dans plusieurs fonds d'archive du web seraient possibles. Les possibilités de navigation qu'offre aujourd'hui l'interface « Wayback Machine » vont dans ce sens : elle « […] permet aux chercheurs de naviguer dans l'archive du web comme ils auraient navigué à l'époque sur le web vivant [et] se double d'une exploration diachronique. » (Bonnel, Oury, 2014, p. 8). La généalogie et les mutations d'un site peuvent ainsi être observées en comparant les différentes versions des captures. (Leetaru 2012)

Dans le cas d'initiatives individuelles d'archivage du web ou provenant d'institutions spécialisées ou de niches, Kalev Leetaru recommande de les rattacher à de plus gros programmes pour ne pas qu'elles disparaissent. Une procédure de soumission pourrait être mise en place qui faciliterait les demandes de ces producteurs « indépendants ». (2012)

6.3.2 Normes de citation

L'usage et la visibilité des archives sont des enjeux qui s'inscrivent au-delà de leur consultation. En effet, si les chercheurs se saisissent petit à petit de ces nouveaux contenus et citent désormais des sources provenant de celles-ci, il s'agit de penser à **normaliser ces citations**. Cette préoccupation participe au travail de leur légitimation, qui ne doit pas échapper aux usages en cours des sources traditionnelles. La mise en

place d'un identifiant unique et permanent de chaque page web archivée participerait à un système de citation efficace dans les publications scientifiques. Comme pour la citation des pages du web vivant, certaines métadonnées comme la date (voire l'heure) de capture de la page sont essentielles pour la constitution de notices complètes. (Leetaru 2012) On observe aujourd'hui une tentative parmi d'autres de standardisation du mode de citation (ici basée sur les standards MLA) impulsé par Internet Archive qui informe les usagers sur la meilleure façon de citer leurs ressources[39]. (Reynolds, 2013, p. 8)

6.3.3 Documenter les robots-crawler

Si les choix documentaires d'acquisition des bibliothécaires sont longtemps restés opaques pour le grand public, il serait envisageable de renverser cette tendance dans le cadre de l'archivage du web. En effet, il serait possible de **documenter les biais (souvent algorithmiques) des crawlers** et autres robots qui moissonnent le web pour l'archiver. De la même façon qu'une transparence des politiques documentaires qui engagent le travail de bibliothécaires autour d'une collecte donnée, la mise en lumière de certains détails techniques propres à un programme peuvent contextualiser telle ou telle collection. Ainsi, comme cela avait été soulevé dans l'étude de cas de la BN, la date d'archivage d'un site peut ne pas correspondre à la date de capture du site. Cette réalité peut constituer un biais majeur pour l'étude d'une chronologie exacte de l'évolution d'un site. Si l'on cherche à comparer, par exemple, le nombre de pages traitant de la candidature à une élection d'un politicien avec celles d'un concurrent, les résultats obtenus ne correspondront pas nécessairement à la réalité du web d'alors. Le nombre d'occurrences peut être un influencé par certaines politiques d'archivage, par l'algorithme selon lequel le robot moissonne le web, etc. (Ben-David, Huurdeman, 2014, p. 107) Il s'agit ici à nouveau d'un biais technique (ou politique) qui se doit d'être éclairé par la documentation des archives du web. Si certains sites ne peuvent pas être collectés intégralement (en raison de liens morts ou de contenus dynamiques difficiles d'approche), il serait utile pour les chercheurs de pouvoir accéder au **« journal » du crawler**, de façon à connaître les lieux où le robot a peut-être buté contre tel ou tel contenu. Les zones blanches des archives peuvent recéler un sens précieux pour ceux qui les étudient.

Par ailleurs, beaucoup de sites dits « dynamiques » adaptent leurs contenus en fonction de l'emplacement physique de l'internaute : dans cette logique, **la géographie du robot** doit être un élément de contexte documenté pour les utilisateurs des

[39] Notamment au sein de la FAQ : http://archive.org/about/faqs.php#265

archives. Elle influe directement sur les contenus affichés (et donc collectés), l'ordre des pages, etc. Un crawler installé en Russie ne collectera pas les mêmes contenus qu'un autre localisé en France, par exemple. (Leetaru 2012)

6.3.4 Archiver le contexte et le web social

En définitive, l'ensemble de ces préoccupations techniques mentionnées plus haut renvoie à la question de **l'archivage du contexte de l'archive**. Nous avons pu mettre en exergue, dans le cas de la BnF, l'archivage du contexte des sites au travers des liens sortants qui donnent à voir l'écosystème global dans lequel le site se déploie. Les métadonnées associées ou la localisation du crawler s'inscrivent dans cette même logique et répondent également aux attentes de certains chercheurs : « […] il s'agit de conserver la trace d'un état antérieur où les contenus sont inséparables de leur « surface d'inscription »[40] : « […] le contenu est très lié à l'architecture [du site] ». » (Chevallier, Illien, 2011, p. 12). De la même manière, l'archivage des documents audiovisuels du web pratiqué par l'INA suppose l'intégration des « paratextes éditoriaux », ces derniers définissent notamment « […] la « grille » d'appréhension à travers laquelle on va regarder [ces documents]. » (Carou, 2007, p. 57). En termes de données contextuelles, les chercheurs attendent spécifiquement « l'URL, la date de capture, la place de la page capturée dans le site, l'arborescence […] et des statistiques de vues ». (Chevallier, Illien, 2011, p. 17). En conservant le contexte, l'archive fait sens et peut faire rayonner tout son pouvoir mémoriel : « […] les différents sites se rattachent à une nébuleuse de sites, et cette nébuleuse il faut en rendre compte, car c'est celle-là qui est en réalité la plus intéressante. » (Chevallier, Illien, 2011, p. 17). Il s'agit également d'un des principes cardinal du théoricien de l'archivistique contemporaine Carol Couture, qui souligne que « Pour l'archiviste, le contexte est cette réalité qui donne tout son sens au contenu des documents d'archives et qui leur permet de remplir leur fonction de preuve et de témoignage. » (Couture, 2000, p.115)

Beaucoup de sites web invitent les visiteurs à interagir avec les contenus présents sur les pages : commenter et partager sont devenues des actions extrêmement courantes. La **dimension sociale** représente aujourd'hui une part substantielle de l'écosystème global d'un site : « This social narrative is an integral part of the content seen by visitors […]. » (Leetaru 2012). C'est tout un environnement de commentaires, entrant en résonnance avec les documents présents sur le site, qu'il faudrait prendre en compte. Sans que de réelles solutions techniques soient véritablement proposées

[40] C'est moi qui souligne.

aujourd'hui, la dimension sociale d'un site se devrait d'être conservée à l'avenir. (Leetaru 2012)

6.3.5 Les archives du web, un agent d'authentification

Au-delà de ses fonctions de préservation de la mémoire du web et de recherche pour la communauté scientifique, les archives du web pourraient constituer, à terme, un **agent d'authentification**. En effet, elles pourraient pointer, par exemple, les changements intervenus sur une page dans un jeu de comparaison entre une page « primaire » (archivée à un moment t) et une page consultée sur le web vivant. Ce travail comparatif prend tout son sens dans le contexte mouvant du web. Les pages des sites gouvernementaux ou médicaux et leurs évolutions pourraient ainsi être authentifiées par les archives. (Leetaru 2012) C'est par ailleurs l'un des objectifs fondamentaux de l'archivistique : garantir, tout comme la fiabilité et l'intégrité, l'authenticité du document, « […] autrement dit qu'il s'agit de sources fiables. » (Duranti 2004).

6.3.6 Conclusion : le cas de Wikipedia et l'effort de sensibilisation

En conclusion, Kalev Leetaru convoque Wikipedia comme le modèle ultime d'une gestion des archives pour le web. En effet, l'encyclopédie libre archive, depuis le début de son existence, toutes les traces des modifications intervenues sur ses pages. En un seul clic, l'internaute peut accéder aux historiques des précédentes versions de chaque page de l'encyclopédie. Elle offre un modèle complètement transparent puisque le code est parfaitement accessible et transposable. Cette gestion des archives d'un site donne à voir ce que pourrait être un système d'archive automatique et normalisé. (2012)

Emily Reynolds insiste, quant à elle, sur l'effort de pédagogie et de sensibilisation à l'archivage du web auprès des étudiants. En impliquant des élèves dans l'élaboration de collections d'archives web, il s'agit de rendre attentives les futures générations à l'importance de ce patrimoine nouveau. (2013, p. 10) A la façon de l'initiative « K-12 Web Archiving »[41], les générations natives numérique peuvent ainsi prendre conscience que les contenus du web ne sont pas éternels et qu'une importante partie de notre mémoire collective se crée, circule et meurt parfois sur la toile.

[41] Pour davantage d'informations sur cette initiative : https://archive-it.org/k12/

7. Conclusion

« Nos mémoires sont de gigantesques prothèses qu'on appelle serveurs, archives ou bibliothèques. »
Michel Melot (Melot 2006)

7.1 Résultats

Comme annoncé dans l'introduction de ce travail, nous avons cherché à dégager les grandes approches et stratégies de collecte de l'archivage du web à l'œuvre aujourd'hui. Nous avons ensuite analysé et comparé deux programmes d'archivage pour saisir plus exactement les processus de travail en vigueur sur le terrain. La question du public des archives du web a ensuite été posée au travers du cas des chercheurs : leurs attentes et besoins ont été décryptés. Enfin, nous avons exposé des horizons innovants possibles pour le futur de ces archives. Afin de mener à bien ces différentes étapes, nous avons établi une revue de la littérature et mené plusieurs entretiens pour combler les lacunes des sources mobilisées.

A l'issue de ce travail, nous sommes parvenus à plusieurs résultats, en filigrane desquels nous retrouvons les préoccupations annoncées dans notre introduction : l'accessibilité, la représentativité, la légitimité, la fiabilité et la destruction des archives du web. Premièrement, les grandes approches de l'archivage ont été dégagées : intégrale, exhaustive, sélective et thématique. Chacune d'entre-elles peut parfois être accompagnée d'une stratégie de collecte particulière : automatisée, semi-automatisée ou manuelle. Nous avons observé, au travers de nos études de cas, à quel point ces différentes approches constituent des cadres théoriques qui se combinent parfois sur le terrain. La littérature invite également à un renouvellement permanent de ces modèles qui doivent s'adapter à de nouvelles réalités de l'archivage. L'analyse comparative des programmes d'archivage de la BN et de la BnF, usant d'approches différentes, illustre bien la nature complémentaire de ces différentes approches qui ne peuvent se déployer que dans un cadre législatif adéquat. Qu'elle s'incarne dans une logique d'échantillonnage ou dans celle d'une collecte large, la question de la représentativité de l'extrême variété des contenus du web préoccupe constamment les administrateurs des programmes.

Nous avons pu observer comment le public des chercheurs considère et appréhende les archives du web. Certaines attentes et résistances ont pu être soulignées : issus d'horizons disciplinaires différents, les chercheurs s'accordent sur la nécessité de conserver une mémoire du web. La disparition des documents numériques natifs inquiète certains d'entre eux. La sélection des contenus à archiver doit être le fruit

d'une politique documentaire aiguisée pour former des collections qui n'apparaissent pas toujours comme légitimes ou fiables aux yeux des chercheurs. Leur mise en place implique également un processus de destruction et de discrimination de certains contenus. Les questions épistémologiques et méthodologiques pour inscrire ces archives dans un usage scientifique établi ne sont pas encore résolues. Enfin, la difficile prise en main des interfaces d'accès des archives et les déplacements contraints au sein des institutions mandataires pour leur consultation découragent ce public.

C'est essentiellement grâce à la production scientifique établie au sein du Consortium IIPC que nous avons pu mettre à jour des innovations et défis futurs de l'archivage du web : ils composent certaines de nos recommandations. Des outils d'analyse du web vivant comme la visualisation des contenus, la recherche au sein de gros ensembles de données ou l'analyse des réseaux sociaux, sont autant de leviers à activer et transposer pour exploiter et mettre en valeur les collections des archives du web. D'autres pistes d'innovations, comme l'archivage des journaux des serveurs pour comprendre l'usage passé des archives, l'exploitation statistique des archives, l'observation de la prolifération d'une idée sur le web au travers des archives, appellent les différents acteurs de l'archivage (archivistes, chercheurs, ingénieurs en informatique) à un travail collaboratif. Les interfaces d'accès aux archives occupent une place majeure dans les projets d'innovation : à la fois vitrines des collections, portes d'accès principales aux contenus et exploratrices de gros volumes de données, elles cristallisent d'importants défis.

Le travail de description des archives et l'inscription systématique de métadonnées sont des recommandations récurrentes des études prospectives sur les archives du web. Plus encore, la description (ou documentation) des robots-crawler peut participer au travail de contextualisation du fonds d'archive ; elle peut permettre d'éviter des biais liés aux algorithmes des robots ou à leur emplacement. Toutes ces pistes concourent à archiver le contexte de l'archive et semblent autant répondre aux attentes de certains chercheurs qu'à inscrire ces nouveaux corpus dans une tradition théorique archivistique. Enfin, l'archivage des données contextuelles est une condition pour qu'à terme, les archives du web puissent être considérées comme un véritable agent d'authentification des contenus numériques. La notion de fiabilité de ces nouvelles archives est ainsi au cœur des préoccupations et représente l'une des conditions nécessaires à leur avènement.

7.2 Limites et perspectives

Au terme de ce travail, nous pouvons observer plusieurs limites à sa réalisation. La première difficulté rencontrée tout au long de notre travail réside dans le peu de sources disponibles sur l'archivage du web. Comme nous l'avons mentionné dans notre méthodologie, c'est principalement sur le web que nous avons trouvé le plus de sources. Nous avons parfois été dérouté par la difficulté technique de certaines d'entre elles : l'archivage du web faisant intervenir plusieurs champs de compétence, certains articles, notamment concernant des innovations futures, se destinaient à un public d'ingénieurs en informatique.

Le changement de granularité d'analyse, à mi-parcours du travail, nous a contraint ä d'abandonner notre objet-test. Nous avons ainsi dû opter pour une analyse plus globale de l'écosystème du site dans son entier. Ce réajustement a notamment rendu caducs la littérature (peu nombreuse) et les entretiens menés avec certains acteurs de la recherche sur les documents pornographiques numériques natifs.

Afin de compléter notre analyse comparative des deux programmes d'archivage, il aurait été intéressant d'ajouter une troisième étude de cas d'un programme du monde anglo-saxon. Cette troisième analyse aurait enrichi les deux premières approches étudiées et permis d'éclairer une communauté linguistique traditionnellement très active dans l'archivage du web. Il aurait été également possible de se concentrer sur un seul programme et de s'attacher à rendre compte de toutes les dimensions propres à celui-ci, en menant notamment une enquête quantitative auprès de son public, par exemple. Enfin, la question budgétaire, nécessairement déterminantes, de ces deux programmes auraient pu être abordée. Nous pouvons en effet émettre l'hypothèse que les ressources financières allouées structurent en grande partie la marge de manœuvre des administrateurs des programmes d'archivage.

Concernant le chapitre sur le public, nous aurions pu, si le temps nous l'avait permis, rencontrer d'autres chercheurs pour étoffer nos sources et les témoignages provenant d'enquêtes déjà réalisées. Il aurait été également intéressant de sélectionner un public de chercheurs précis – historiens ou sociologues, par exemple – et de procéder à une enquête spécifique auprès de cette communauté scientifique. Loin de considérer le public comme un bloc monolithique, nous aurions pu choisir d'autres segments que celui des chercheurs et observer comment les besoins et attentes varient suivant les publics. L'organisation de focus groupes qui discuteraient ensemble de leurs attentes et représentations des archives auraient, par exemple, permis de repérer les affinités et les alliances d'un public spécifique. Par ailleurs, il aurait été intéressant de

confronter les différents responsables des programmes étudiés aux innovations et recommandations proposées dans notre travail ; parmi elles, quelles seraient les plus applicables aujourd'hui pour la BN ou la BnF ?

Afin d'éviter le spectre de l' « Erreur HTTP 404 »[42] à laquelle nous avons toutes et tous été confrontés, l'établissement d'une mémoire numérique apparaît si ce n'est comme urgent, du moins légitime. Certains chercheurs l'ont bien compris et n'hésitent pas, comme Kalev Leetaru, à déclarer dans leurs travaux : « In the web era, we are repeating this cycle of loss, not through a fire or other sudden even that destroyed the Library of Alexandria, but rather thourgh inaction : we are simply not collecting it. » (2012). La lutte contre la disparition des contenus nativement numériques ne vise pas l'exhaustivité : comme toute archive, celles du web demeurent fragmentées car choisies, puis architecturées au sein de corpus. Les institutions et les acteurs en charge de la collecte de ce patrimoine mondial pensent le gigantisme de cette masse documentaire pour en offrir les échantillons les plus remarquables et agencés de la meilleure façon qui soit. Les quelques exemples, extrêmement enthousiasmants, de l'usage de ces nouvelles archives que nous avons pu souligner dans ce travail montrent toute la richesse des traces circulant sur le web.

En concentrant un maximum les actions du quotidien d'une société sur son réseau, Internet tend à devenir un lieu de notre histoire mondiale. La trace, le signe ou l'indice numérique nous invite à considérer le web et son archivage comme une véritable archéologie des pratiques humaines. Comme le pressentait un chercheur interrogé dans l'étude menée par la BnF, il ne s'agira plus, pour les historiens et archéologues de demain, d'investir seulement les strates des sols à la découverte des vestiges du temps, mais désormais aussi celles du web, à la recherche de notre identité profonde.

[42] Cette erreur du protocole de communication HTTP sur Internet indique à l'utilisateur que le contenu désiré n'existe pas ou plus.

Bibliographie

ABBOU, Julie, 2013. Calenda le calendrier des lettres et sciences humaines et sociales. *Calenda.org* [en ligne]. 6 décembre 2013. [Consulté le 21.06.2015]. Disponible à l'adresse : http://calenda.org/267910

AUBRY, Sara et al., 2008. Méthodes techniques et outils. *Documentaliste-Sciences de l'Information* [en ligne]. Avril 2008. Vol. 45. p.12-20. [Consulté le 11.05.2015]. Disponible à l'adresse : http://www.cairn.info/revue-documentaliste-sciences-de-l-information-2008-4-p-12.htm

BALZARDI, Elena, 2008. Le projet e-Helvetica de la Bibliothèque nationale suisse. *Admin.ch* [en ligne]. 14 décembre 2006. 28 février 2008. [Consulté le 15.02.2015]. Disponible à l'adresse : http://www.nb.admin.ch/nb_professionnel/01693/01699/01873/01893/index.html?lang=fr

BEAUDOUX, Clara, 2013. Archiver le web, un devoir de mémoire GIGAntesque. *France Info* [en ligne]. 5 avril 2013. [Consulté le 21.06.2015]. Disponible à l'adresse : http://www.franceinfo.fr/vie-quotidienne/high-tech/article/archiver-le-web-un-devoir-de-memoire-gigantesque-242673

BEN-DAVID, Anat, HUURDEMAN, Hugo, 2014. Web Archive Search as Research : Methodological and Theoretical Implications. *Alexandria*. 2014. Vol. 25. No 1. [Consulté le 10.06.2015]. Disponible en téléchargement gratuit à l'adresse : http://www.academia.edu/8993065/Web_archive_search_as_research_Methodological_and_theoretical_implications

BLOCH, Marc, 1999. *Apologie pour l'histoire ou métier d'historien*. Paris : Colin, 1999. Consulté en ligne à l'adresse : http://classiques.uqac.ca/classiques/bloch_marc/apologie_histoire/bloch_apologie.pdf

BN, 2012a. Sites web – Archives Web Suisse. *Admin.ch* [en ligne]. 12 décembre 2012. [Consulté le 02.05.2015]. Disponible à l'adresse : http://www.nb.admin.ch/nb_professionnel/01693/01695/01705/index.html?lang=fr

BN, 2012b. e-Helvetica. *Admin.ch* [en ligne]. 25 juillet 2012. [Consulté le 25.02.2015]. Disponible à l'adresse : http://www.nb.admin.ch/nb_professionnel/01693/index.html?lang=fr

BN, 2011. Dépôt légal. *Admin.ch* [en ligne]. 24 janvier 2011. [Consulté le 08.05.2015]. Disponible à l'adresse : http://www.nb.admin.ch/dienstleistungen/swissinfodesk/03034/03232/03702/?lang=fr

BNF, 2015a. Collectes ciblées de l'internet français. *Bnf.fr* [en ligne]. 26 mars 2015. [Consulté le 08.04.2015]. Disponible à l'adresse : http://www.bnf.fr/fr/collections_et_services/anx_pres/a.collectes_ciblees_arch_internet.html

BNF, 2015b. Application pour le dépôt légal du web. *Bnf.fr* [en ligne]. 8 avril 2015. [Consulté le 08.04.2015]. Disponible à l'adresse : http://www.bnf.fr/fr/professionnels/dlweb_boite_outils/a.dlweb_applications.html

BNF, 2014. Dépôt légal des sites web. *Bnf.fr* [en ligne]. 4 décembre 2014. [Consulté le 12.03.2015]. Disponible à l'adresse : http://www.bnf.fr/fr/professionnels/depot_legal/a.dl_sites_web_mod.html

BNF, 2014b. Les signets de la Bnf : Wayback machine. *Les signets de la Bnf* [en ligne]. 6 août 2014. [Consulté le 30.05.2015]. Disponible à l'adresse : http://signets.bnf.fr/html/notices/n_3579.html

BNF, 2014c. Qu'est-ce que le dépôt légal ? *Bnf.fr* [en ligne]. 1er octobre 2014. [Consulté le 11.06.2015]. Disponible à l'adresse : http://www.bnf.fr/fr/professionnels/depot_legal_definition/s.depot_legal_organisation.html?first_Art=non

BNF, 2006. Dossier de presse : Internet en campagne. *Bnf.fr* [en ligne]. 25 octobre 2006. [Consulté le 06.05.2015]. Disponible à l'adresse : http://www.bnf.fr/documents/dp_internet_campagne.pdf

BONNEL, Sylvie, OURY, Clément, 2014. La sélection de sites web dans une bibliothèque nationale encyclopédique : une politique documentaire partagée pour le dépôt légal de l'internet à la BnF. IFLA World Library and Information Congress 80th IFLA General Conference and Assembly, Lyon, 16-22 August 2014 [en ligne]. [Consulté le 04.04.2015]. Disponible à l'adresse : http://library.ifla.org/998/1/107-bonnel-fr.pdf

CAROU, Alain, 2007. Archiver la vidéo sur le web : des documents ? Quels documents ?. *Bulletin des bibliothèques de France* [en ligne]. 2007. N°2. [Consulté le 22.04.2015]. Disponible à l'adresse : http://bbf.enssib.fr/consulter/bbf-2007-02-0056-012

CHAIMBAULT, Thomas, 2008. L'archivage du web [en ligne]. Dossier documentaire. Villeurbanne : enssib. 2008. [Consulté le 02.03.2015]. Disponible à l'adresse : http://www.enssib.fr/bibliotheque-numerique/documents/1730-l-archivage-du-web.pdf

CHEVALLIER, Philippe, ILLIEN, Gildas, 2011. Les Archives de l'Internet : une étude prospective sur les représentations et les attentes des utilisateurs potentiels [en ligne]. Bibliothèque nationale de France. 2011. [Consulté le 01.05.2015]. Disponible à l'adresse : http://www.bnf.fr/documents/enquete_archives_web.pdf

COUTURE, Carol, 2000. *Les fonctions de l'archivistique contemporaine*. Presses de l'Université du Québec, 2000.

DURANTI, Luciana, CHABIN, Marie-Anne, 2004. La conservation à long terme des documents dynamiques et interactifs : InterPARES 2. *Document numérique* 2/2004 [en ligne], vol. 8, 2004. [Consulté le 20.06.2015]. Disponible à l'adresse : http://www.cairn.info/revue-document-numerique-2004-2-page-73.htm#anchor_citation

FRANCE, 2015. *Code du patrimoine, Partie législative, Dépôt légal, art. L131-2* [en ligne]. Legivrance. 24 juillet 2009. [Consulté le 22.03.2015]. Disponible à l'adresse : http://www.legifrance.gouv.fr/affichCode.do;jsessionid=77BCE731A86D0A1C02D19877FC37F3C7.tpdila20v_2?idSectionTA=LEGISCTA000006159934&cidTexte=LEGITEXT000006074236&dateTexte=20150429

GENIN, Christine, 2012. Archiver l'hypertexte. *Revue de la bibliothèque nationale de France* [en ligne]. Mars 2012. N°42. 96 p. [Consulté le 17.05.2015]. Disponible à l'adresse : http://www.cairn.info/revue-de-la-bibliotheque-nationale-de-france-2012-3-page-21.htm [accès par abonnement]

GHARSALLAH, Mehdi, 2004. Archivage du web français et dépôt légal des publications électroniques. *Documentaliste – Sciences de l'Information* [en ligne]. 2004. [Consulté le 02.06.2015]. Disponible à l'adresse : http://archivesic.ccsd.cnrs.fr/sic_00001311/fr/

GREFFET, Fabienne, 2012. Le web dans la recherche en science politique [en ligne]. *Revue de la Bibliothèque nationale de France* [en ligne], n°40. 2012. [Consulté le 06.04.2015]. Disponible à l'adresse : www.cairn.info/load_pdf.php?ID_ARTICLE=RBNF_040_0078

HAN, Myung-Ja, 2012. Les niveaux de description des métadonnées : un nouveau défi pour les bibliothécaires experts de catalogage et de métadonnées. IFLA World Library

and Information Congress 78th IFLA General Conference and Assembly, Helsinki [en ligne]. [Consulté le 27.05.2015]. Disponible à l'adresse : http://conference.ifla.org/past-wlic/2012/80-han-fr.pdf

ILLIEN, Gildas, 2011. Une histoire politique de l'archivage du web. *Bulletin des bibliothèques de France* [en ligne], n°2, 2011. [Consulté le 23.05.2015]. Disponible à l'adresse : http://bbf.enssib.fr/consulter/bbf-2011-02-0060-012

ILLIEN, Gildas, 2008. Le dépôt légal de l'internet en pratique. *Bulletin des bibliothèques de France* [en ligne], n° 6, 2008. [Consulté le 05 mai 2015]. Disponible à l'adresse : http://bbf.enssib.fr/consulter/bbf-2008-06-0020-004

JOUTARD, Philippe, 2013. Révolution numérique et rapport au passé. *Le Débat* [en ligne], n°177, 2013. [Consulté le 30.05.2015]. Disponible à l'adresse : http://www.cairn.info/revue-le-debat-2013-5-page-145.htm

LEETARU, Kalev H., 2012. A vision of the role and future of web archives. IIPC 2012 General Assembly, [en ligne], 2012. [Consulté le 15.04.2015]. Disponible à l'adresse : http://netpreserve.org/sites/default/files/resources/VisionRoles.pdf

LOCHER, Hansueli, 2015. Archives Web Suisse – Notice Archivage. *Admin.ch* [en ligne]. 30 janvier 2015. [Consulté le 04.05.2015]. Disponible à l'adresse : http://www.nb.admin.ch/nb_professionnel/01693/01699/01873/01895/index.html?lang=fr

LUPOVICI, Catherine, ILLIEN, Gildas, AUBRY, Sara, OURY, Clément, LASFARGUES, France, HAFRI, Younès, WENDLAND, Bert, 2006. Web Archiving at BnF. *IIPC netpreserve.org* [en ligne]. Septembre 2006. [Consulté le 02.04.2015]. Disponible à l'adresse : http://www.netpreserve.org/sites/default/files/resources/BnFnews200609.pdf

MELOT, Michel, 2006. *La sagesse du bibliothécaire*. Editions Jean-Claude Béhar. Sagesse d'un métier.

MERZEAU, Louise, 2003. Web en stock. *Cahier de médiologie* [en ligne]. 2003. P. 158-167. [Consulté le 05.06.2015]. Disponible à l'adresse : https://halshs.archives-ouvertes.fr/halshs-00487319/document

MEYER, Eric T., THOMAS, Arthur, SCHROEDER, Ralph, 2011. Web Archives : The Future(s). *IIPC netpreserve.org* [en ligne]. 2011. [Consulté le 12.04.2015]. Disponible à l'adresse : http://netpreserve.org/sites/default/files/resources/2011_06_IIPC_WebArchives-TheFutures.pdf

MUSSOU, Claude. Et le web devint archive : enjeux et défis. *Ina-expert.com* [en ligne]. Juin 2012. [Consulté le 14.03.2015]. Disponible à l'adresse : http://www.ina-expert.com/e-dossier-de-l-audiovisuel-sciences-humaines-et-sociales-et-patrimoine-numerique/et-le-web-devint-archive-enjeux-et-defis.html

OURY, Clément, 2012. Archivage du web : BigData & PétaBox : Le dépôt légal du web : BnF. *Labo.bnf* [en ligne]. 17 octobre 2012. [Consulté le 23.04.2015]. Disponible à l'adresse : http://labo.bnf.fr/video_121017_atelier_dlweb_2.html

PEYSSARD, Jean-Christophe, GINOUVES, Véronique, 2012. Internet Archive. *Aldebaran.revues.org* [en ligne]. 2 septembre 2012. [Consulté le 02.06.2015]. Disponible à l'adresse : http://aldebaran.revues.org/6339

REYNOLDS, Emily, 2013. Web Archiving Use Cases. Library of Congress, UMSI, ASB13 [en ligne]. Mars 2013. [Consulté le 18.05.2015]. Disponible à l'adresse : http://netpreserve.org/sites/default/files/resources/UseCases_Final_1.pdf

SALAUN, Jean-Michel, 2007. « La redocumentarisation, un défi pour les sciences de l'information », *Etudes de communication* [en ligne], 30. 2007. [Consulté le 24.05.2015]. Disponible à l'adresse : http://edc.revues.org/428

SAMPLE, Ian, 2015. Google boss warns of 'forgotten century with email and photos at risk. *The Guardian* [en ligne]. 13 février 2015. [Consulté le 09.03.2015]. Disponible à l'adresse : http://www.theguardian.com/technology/2015/feb/13/google-boss-warns-forgotten-century-email-photos-vint-cerfy

SIGNORI, Barbara, 2015a. *AW : Demande entretien HEG-ID / Travail de bachelor*. [message électronique]. 6 mai 2012.

SIGNORI, Barbara, 2015b. Archives Web Suisse – Bases. *Admin.ch* [en ligne]. 5 janvier 2015. [Consulté le 17.03.2015]. Disponible à l'adresse : http://www.nb.admin.ch/nb_professionnel/01693/01695/01705/index.html?lang=fr

SIGNORI, Barbara, 2011. Archives Web Suisse – Notice Collecte. *Admin.ch* [en ligne]. 15 janvier 2011. [Consulté le 03.02.2015]. Disponible à l'adresse : http://www.nb.admin.ch/nb_professionnel/01693/01699/01873/01895/index.html?lang=fr

SUISSE, 1992. *Loi fédérale sur la Bibliothèque nationale suisse (LBNS) du 18 décembre 1992 (Etat le 1er janvier 2012)* [en ligne]. 18 décembre 1992. RS 432.21. [Consulté le 14.04.2015]. Disponible à l'adresse : https://www.admin.ch/opc/fr/classified-compilation/19920349/

Annexes

Entretien avec Barbara Signori, responsable du programme e-Helvetica, reçu le 6 mai 2015. (Courriel)

Questionnaire e-Helvetica – Archives Web Suisse

Processus de travail

1. Comme évoqué par C. Couture (dans son ouvrage « Les fonctions de l'archivistique contemporaine ») et d'autres archivistes, un document devient signifiant dans, par et au travers de son contexte. Dans cette perspective, comment la BN se positionne-t-elle face au potentiel archivage des interactions d'hyperliens ? Plus largement, comment rend-elle compte des réalités du web 2.0 (dans l'archivage des commentaires d'usagers, par exemple)?

Nous collectons tous les contenus d'un site web, ainsi que les commentaires, ceci pour autant que cela soit techniquement possible et que la grandeur maximale pour l'archivage ne soit pas dépassée.

2. La BN a une approche strictement sélective et thématique de l'archivage du web. Quelles sont, selon vous, les qualités et les forces d'une telle approche ?
A l'inverse, quels en sont les défauts/limites majeurs ?
Que pensez-vous des approches intégrale et/ou exhaustive ?

La sélection garantit une certaine qualité des sites et nous avons connaissance de ce que nous avons enregistré dans les Archives Web. Dans le cas d'un moissonnage de domaine, c'est la quantité qui est décisive. Les deux variantes se valent, ce qui est déterminant, ce sont les conditions de base.

3. Comment conjuguer processus sélectif des collections et caractère représentatif ?

Nos partenaires procèdent à la sélection et nous essayons ensemble de construire une collection représentative.

4. Vous arrive-t-il, et si oui à quelle fréquence, d'être empêchée par un particulier qui ne souhaite pas que son site web soit archivé ? Quelle est votre stratégie dans ce cas-là ?

Oui, cela arrive, mais à un très petit pourcentage. Nous respectons la décision et nous ne collectons et n'archivons pas le site.

5. En ce qui concerne la marge de manœuvre laissée aux cantons dans le choix des sites Internet à archiver, celui-ci peut-il être discuté (voire invalidé) par la BN (en dehors d'impossibilités techniques à archiver le site en question) ?

Les directives de collecte ont été définies en commun avec les bibliothèques cantonales. Si nécessaire, elles sont aussi révisées en commun.

6. La BN rend accessible sa collection « Archives Web Suisse » au sein de ses locaux : pensez-vous qu'il serait positif qu'elle soit également accessible depuis les différentes institutions cantonales ?

L'accès est aussi possible dans les locaux des bibliothèques partenaires, pour autant qu'elles aient effectué les installations nécessaires.

BN et BnF

7. Pourquoi avoir renoncé au moissonnage de tout le domaine .ch ? (injonctions financières ou philosophie archivistique ?)
Si cela était amené à changer, quid des contenus alternatifs (et volatiles), comme les documents pornographiques ?
Que pensez-vous de l'approche de la BnF qui cherche au contraire à moissonner l'entier de son domaine, cherchant à rendre compte d'un aperçu le plus exhaustif possible de son web à un moment donné ?

Il nous manque les bases légales pour le faire. Le dépôt légal de la BnF inclut aussi les contenus du web.

8. La BnF a une politique d'archivage du web en tout point opposée à celle de la BN : comment la considérez-vous ?

Nous avons des conditions de base différentes. De ce point de vue, les approches ne peuvent pas être comparées.

9. Le cadre légal suisse ne prévoit pas l'obligation pour les éditeurs de laisser leurs sites à archiver. Que pensez-vous du dépôt légal numérique français qui lui autorise les institutions mandataires à copier les sites web sans solliciter l'autorisation préalable des éditeurs? S'agirait-il de changer la législation suisse ?

L'archivage web se trouve simplifié en termes d'obtention des droits. Lorsque la demande de collecte tombe, on économise des ressources.

Périmètre de collecte

10. Toute une partie du web est évacuée du programme « Archives Web Suisse », notamment le « web invisible » (ou « web profond ») : considérez-vous ces limites techniques comme un frein à la qualité « patrimoniale » de la collection ?

C'est certainement un frein. Nous pouvons collecter uniquement ce qui peut être collecté avec la technologie actuelle. La technologie est en constante évolution.

11. Il arrive que certains sites compatibles avec le périmètre de la collection ne puissent être archivés pour des raisons techniques. Dans ce cas, que faites-vous de ce site ? Intègre-t-il une « liste d'attente » ?

Oui

12. Les sites web au contenu pornographique sont exclus au même titre, par exemple, que des sites web au contenu raciste : comment expliquez-vous cette exclusion ?

Comment considérez-vous la dimension patrimoniale et politique de la production pornographique ?

Cette exclusion est une décision juridique.

13. Pourquoi est-ce que le périmètre de la collecte exclue-t-elle les blogs ? Pensez-vous à moyen terme intégrer ce type de site web ?

Là aussi les raisons juridiques sont déterminantes. Les blogs seront autorisés prochainement.

14. Le listing des sites qui ne sont pas collectionnés n'est pas exhaustif et amené à changer : à votre avis, quelles en seront les évolutions futures ?

Les formes de publications du Web sont toujours plus innovatrices. Le plus souvent nous échouerons devant des obstacles techniques.

Volatilité des éphémères

15. Parmi les contraintes techniques récurrentes, on note celle de l'utilisation de Flash (et JavaScript). Comment la BN se positionne-t-elle face à la prolifération des contenus dynamiques et à leur intégration dans les collections ? (Que faire des documents numériques natifs associés à un site ?)

Lorsque nous ne pouvons pas collecter un site pour des raisons techniques, ou lorsque nous ne pouvons pas le collecter de manière suffisante, nous ne l'archivons pas.
Lorsque nous ne pouvons pas afficher un site que nous avons cependant réussi à collecter complètement, nous l'archivons.

16. Comment la BN procède-t-elle spécifiquement pour s'emparer des documents dynamiques présents au sein d'un site web qu'elle souhaite archiver ?
Si les contenus audiovisuels du site sont amenés à se renouveler souvent, les intègre-t-elle au moins une fois par année ?

La plupart des sites sont collectés et archivés une fois par année.

La BN à l'international

17. La Suisse est membre de l'IIPC : quelle est son implication dans cet organisme ?

Nous sommes membres du Comité de pilotage (Steering Committee) et nous sommes représentés dans les Groupes de travail (Working Groups).

18. Explicitez les collaborations internationales essentielles dans le cadre de la collection « Archives Web Suisse ».

Nous utilisons les outils qui sont développés à l'IIPC et nous prenons part aux projets de collectes en commun.

Futurs et perspectives

19. A titre d'exemple, l'une des pistes d'innovation autour de l'utilisation des archives du web réside dans l'étude de la naissance des idées sur le web et la façon dont elles se propagent. La profondeur de l'archivage appliquée par la BN dans son programme permettrait peut-être une application allant dans ce sens. Est-ce que vous réfléchissez à de nouvelles façons d'utiliser la collection et si oui, comment ?

Les réflexions sur l'utilisation des Archives Web Suisse ne sont pas encore achevées.

20. L'un des constats récurrents des études sur les programmes d'archivage du web est l'absence d'intégration des chercheurs dans le processus de collecte et la mise en place des interfaces de consultation. En tant que responsable, que pensez-vous de ce constat ?

Je trouve qu'il est important d'impliquer les utilisateurs pour ce qui concerne l'accès aux documents.

21. En quoi est-ce que l'archivage du web et plus largement les questions de la mémoire numérique vous semblent importantes aujourd'hui ?

Il y a beaucoup de défis dans ce domaine. La collaboration et la coordination - nationales et internationales - sont importantes, ceci afin de ne pas à chaque fois tout recommencer depuis le début et faire la même chose.

22. Les résistances à ce type d'initiative sont multiples : comment sensibiliseriez-vous de nouveaux partenaires potentiels ?

Nous leur démontrons les avantages de la collaboration.